THE ANNALS
of
MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE
OF MATHEMATICAL STATISTICS

Contents

PAGE

The Cyclic Effects of Linear Graduations Persisting in the Differences of the Graduated Values. EDWARD L. DODD 127

On the Distribution of Wilks' Statistic for Testing the Independence of Several Groups of Variates. A. WALD and R. J. BROOKNER. 137

The Mean Square Successive Difference. J. VON NEUMANN, R. H. KENT, H. R. BELLINSON, and B. I. HART 153

The Return Period of Flood Flows. E. J. GUMBEL 163

On the Foundations of Probability and Statistics. R. VON MISES. 191

Probability as Measure. J. L. DOOB 206

Discussion of Papers on Probability Theory. R. VON MISES and J. L. DOOB 215

Continued Fractions for the Incomplete Beta Function. LEO A. AROIAN 218

Notes:

- Note on the Distribution of Non-central t with an Application. CECIL C. CRAIG 224
- Note on an Application of Runs to Quality Control Charts. FREDERICK MOSTELLER 228
- Test of Homogeneity for Normal Populations. G. A. BAKER 233
- A Note on the Power of the Sign Test. W. MAC STEWART 236
- Moments of the Ratio of the Mean Square Successive Difference to the Mean Square Difference in Samples from a Normal Universe. J. D. WILLIAMS 239

Vol. XII, No. 2 — June, 1941

THE ANNALS OF MATHEMATICAL STATISTICS

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore, Md., or to the Secretary of the Institute of Mathematical Statistics, E. G. Olds, Carnegie Institute of Technology, Pittsburgh, Pa.

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $4.00 per year. Single copies $1.25. Back numbers are available at the following rates:

Vols. I-IV $5.00 each. Single numbers $1.50.
Vols. V to date $4.00 each. Single numbers $1.25.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879

THE CYCLIC EFFECTS OF LINEAR GRADUATIONS PERSISTING IN THE DIFFERENCES OF THE GRADUATED VALUES

By Edward L. Dodd

University of Texas

1. Scope of inquiry. Slutzky [1] applied the moving sum, the repeated moving sum, and other linear processes to random numbers obtained from lottery drawings. But the graph of the *moving sum* becomes, when the vertical scale is changed in the ratio of n to 1, the graph of the *moving average*, the simplest form of *graduation*. When cyclic effects are studied, there is no essential difference between a moving sum and a moving average, nor between a general linear process with coefficients $a_1, a_2, \cdots, a_s$, having sum $A \neq 0$ and the corresponding *graduation*, with coefficients $a_i' = a_i/A$. Thus Slutzky's work throws considerable light upon graduation, although his main interest was in summation.

Slutzky found that the graphs of moving sums of random numbers bore strong resemblance to graphs of economic phenomena, such as [1, p. 110] that of English business cycles from 1855 to 1877. In fact, Slutzky regards the fluctuations in economic phenomena as due largely to a synthesizing of random causes.

In general the undulatory character of such values cannot be described as periodic; since the waves are of different length. But Slutzky found that, upon operating on random data having mean zero and constant variance, the resulting values approach a sinusoidal limit under certain conditions,—in particular, when a set of n summations by twos is followed by m differencings, and as $n \to \infty$, $m/n \to a$ constant. Romanovsky [2] generalized this result by taking successive summations of s consecutive elements of the data, with $s \geqq 2$; but required that $m/n \to \alpha \neq 1$. However, the cases which are of interest to me just now are those for which $m = n - 1$ or $m = n - 2$; and for these cases $m/n \to 1$. Romanovsky considers the case of $m = n - 1$,—not, however, as leading to a sinusoidal limit,—and gives in formula (46) the value of a coefficient of correlation—which I deduce directly. From his formula (43) a corresponding coefficient of correlation can be obtained for the case of $m = n - 2$, as the sum of certain products. A more simple expression than this I need, which I obtain directly. In my treatment, these coefficients are the cosines of angles; and the ratio of such an angle to a whole revolution is an expected frequency of occurrence.

After setting forth in Section 2 some preliminary formulas, I treat in Section 3 the results of applying to random data an indefinite number $k + 2$ of summations or averagings, followed by k differencings—the number of terms in a sum remaining fixed. In Section 4, however, only a few differencings are applied to a

graduation. In particular the Spencer 21-term formula is studied in some detail. In former papers [3, 4] I have dealt with the immediate effects of graduations upon random data.

The question to be considered in this paper is this: *Do the cyclic effects appearing in the graduated values persist in the successive differences? And, if so, do these affects fade out gradually or on the other hand, do they come to a rather abrupt termination?*

These differences of graduated values, indeed, up to the third, fourth or fifth are of considerable importance. Henderson [5] defines the smoothing coefficient of a given graduation as the ratio of the theoretical standard deviation of the third differences for the graduated values to that for the original values or data.

2. Preliminary notions and formulas. The data to be graduated will be supposed to be independent, or uncorrelated, or as Slutzky expresses it, "incoherent." This will imply that the expected value of the product of two different chance variates is the product of their expected values.

Now the operations of summing and differencing as used here are not inverse. To illustrate: Given as independent $u, v, w, x, y, z, \cdots$. Summing by twos yields the sequence $u + v, v + w, w + x, x + y, y + z, \cdots$. But the first differences of these numbers, $w - u, x - v, y - w, z - x, \cdots$ are alternately correlated, thus $w - u$ is negatively correlated with $y - w$; $x - v$ with $z - x$, etc. Indeed, successive differencing following successive summing does not lead back to the original condition of incoherency. However, under certain conditions, the resulting coherency may be so slight that the final succession of numbers may have just about the same chaotic properties as the succession of data.

In my paper [3, p. 262], I set forth a number of features on the basis of which a cycle length could be defined. One of these involves the frequency of maxima. Given independent chance variables, each subject to the same law of distribution,

$$(1) \qquad P(x_i \leqq x) = \Phi(x);$$

where $\Phi(x)$ has a derivative $\phi(x)$. It is then easy to see that the expected relative frequency of maxima is 1/3. That is:

$$(2) \qquad P(x_{i-1} \leqq x_i \geqq x_{i+1}) = \int_{-\infty}^{\infty} [\Phi(x)]^2 \phi(x)\, dx = 1/3.$$

Now, for a given feature, a cycle length is defined as the reciprocal of the theoretic relative frequency. Then the cycle length here for *maxima* is three. It is well known that averaging tends to remove maxima. Thus, upon averaging or summing, the cycle length tends to increase. It is almost as well known that differencing tends to increase the frequency of maxima, and thus decrease cycle length. For if $z_i = \Delta y_i = y_{i+1} - y_i$, then between two maxima of y_i, there is at least one minimum (strong and weak) of y_i; and following this minimum and before passing the next maximum of y_i there is *at least one* maximum of z_i. Successive differencing tends to reduce the cycle length of maxima from 3 to 2,

that is to make the graph a perfect zig-zag where positive and negative values of z_i alternate. A set of differencings following a set of summings may bring the cycle length from some fairly large number back to about 3, and thus restore something like the original chaotic appearance in the graph.

In dealing with the foregoing $\Phi(x)$ or $\phi(x)$ in (2), it was not assumed that the distribution be normal. But, in what follows, it will be assumed that

$$\phi(x) = \frac{1}{\sigma(2\pi)^{1/2}} e^{-(x-\mu)^2/2\sigma^2}; \tag{3}$$

and, for convenience, μ will be taken as zero—that is, the data will be supposed given as deviations from their theoretic mean. Actually, the data used by Slutzky and the data I have used belong to a rectangular distribution, as noted in my former paper. Nevertheless the close agreement between actual and expected results seems to indicate [3, p. 263] that the theory is in general applicable. It is well known that averaging of observations from non-normal distributions may lead rather quickly to an approximately normal distribution.

Given n real numbers, $a_1, a_2, \cdots, a_n$, let

$$y_j = a_1x_i + a_2x_{i+1} + \cdots + a_nx_{i+n-1}; \qquad i = 1, 2, 3, \cdots. \tag{4}$$

Then y_j is the moving *sum* if each $a_r = 1$. Slutzky takes $j = i$ or $j = i + n - 1$. Again, y_j is the moving *average* if each $a_r = 1/n$. For graduation in general, the condition $\Sigma a_r = 1$ is imposed; and usually $j = i + (n + 1)/2$. If n is odd, y_j is thus associated with the middle x.

Under the assumption that the x's are independent and normally distributed about mean zero, with constant variance, I have proven [3, p. 256]: The probability that for any specified j, $y_{j-1} < 0$, and $y_j > 0$ is given by $P = \theta/360°$, where

$$\cos\theta = \sum_{r=1}^{n-1} a_r a_{r+1} \Big/ \sum_{r=1}^{r=n} a_r^2. \tag{5}$$

The expected relative frequency of up-crossings of the graph of the y's through the zero base line is then $\theta/360°$. That is: $\theta/360°$ is the expected relative frequency of a change in the sign of y from $-$ to $+$; also, of a change in sign from $+$ to $-$.

But, as $\Delta y_j = y_{j+1} - y_j$, it follows that

$$\Delta y_j = b_1x_i + b_2x_{i+1} + \cdots + b_nx_{i+n-1} + b_{n+1}x_{i+n}, \tag{6}$$

where

$$b_1 = -a_1, \quad b_{n+1} = a_n, \quad b_r = a_{r-1} - a_r, \qquad r = 2, 3, \cdots, n-1 \tag{7}$$

and since a maximum for the y's at y_i occurs when $\Delta y_{i-1} > 0$, $\Delta y_i < 0$, it follows that the theoretic frequency therefor is $\theta'/360°$, where

$$\cos\theta' = \sum_{r=1}^{n} b_r b_{r+1} \Big/ \sum_{r=1}^{n+1} b_r^2. \tag{8}$$

In a similar manner, by using *second* differences, we get the expected relative frequency $\theta''/360°$ for inflexional points, in specified direction. Moreover, $\theta \leqq \theta' \leqq \theta'' \leqq \cdots \leqq 180°$; since inflections must be at least as frequent as maxima, etc.

If the foregoing formulas are applied to the identical "graduation" $y_j = x_j$, then $\cos \theta = 0$, $\cos \theta' = -1/2$, $\cos \theta'' = -2/3$. In fact,

$$\cos \theta^{(t)} = -t/(t+1). \tag{9}$$

This follows from the fact that the b's and similar coefficients are the binomial coefficients; and

$$\sum_{r=0}^{t} {}_tC_r^2 = {}_{2t}C_t; \qquad \sum_{r=0}^{t-1} {}_tC_r \cdot {}_tC_{r+1} = {}_{2t}C_{t-1}. \tag{10}$$

Thus repeated differencing leads toward the perfect zig-zag. An extension of this feature will be taken up in the next section.

3. Repeated summing and differencing. To indicate the *result* of the summing of n consecutive numbers in a sequence, I shall use the notation 1^n. And the difference $\Delta y_j = y_{j+1} - y_j$ will be indicated by $-1, 0^{n-1}, 1$. Thus if $n = 3$, 1^3 and $-1, 0^2, 1$ will stand respectively for

$$y_i = x_{i-1} + x_i + x_{i+1}; \qquad \Delta y_i = -x_{i-1} + 0x_i + 0x_{i+1} + x_{i+2}. \tag{11}$$

If, now, $z_i = y_{i-1} + y_i + y_{i+1}$, then

$$z_i = x_{i-2} + 2x_{i-1} + 3x_i + 2x_{i+1} + x_{i+2}. \tag{12}$$

Since (n) is often used to indicate the *operation* of summing n consecutive numbers, we may write

$$(3)^2 = 1, 2, 3, 2, 1; \qquad (n)^2 = 1, 2, \cdots, (n-1), n, (n-1), \cdots, 2, 1. \tag{13}$$

Then, for $n \geq 2$,

$$\Delta(n)^2 = -1^n, 1^n; \qquad \Delta^2(n)^2 = 1, 0^{n-1}, -2, 0^{n-1}, 1. \tag{14}$$

And, since the operations of summing and differencing are commutative, we are lead to

$$F_n^k = (-1)^k \Delta^k (n)^k = {}_kC_0, 0^{n-1}, -{}_kC_1, 0^{n-1}, {}_kC_2, 0^{n-1}, \cdots, (-1)^k {}_kC_k; \tag{15}$$

as may be established by induction. For from the foregoing, it follows that

$$(-1)^k \Delta^k (n)^{k+1} = {}_kC_0^n, -{}_kC_1^n, \cdots, (-1)^k {}_kC_k^n. \tag{16}$$

Then, since ${}_{k+1}C_r = {}_kC_r + {}_kC_{r-1}$, we conclude that

$$F_n^{k+1} = (-1)^{k+1}(n)^{k+1} = {}_{k+1}C_0^n, 0^{n-1}, -{}_{k+1}C_1^n, 0^{n-1}, \cdots, (-1)^{k+1} {}_{k+1}C_{k+1}^n. \tag{17}$$

If now $n \geqq 2$, then from (5) and (15) we find that

$$\cos \theta = 0; \qquad \theta/360° = 1/4. \tag{18}$$

Thus, the expected frequency of the changes in sign of $\Delta^k(n)^k$ is the same as that for the raw or ungraduated data. Moreover, if $n \geqq 3$, (8) leads to $\cos \theta' = -1/2$, found for the data. For, in this case, at least two zero coefficients intervene between any two non-zero coefficients. And thus

$$(19) \qquad \cos \theta' = -\sum_{r=0}^{k} {}_kC_r^2 \Big/ 2 \sum_{r=0}^{k} {}_kC_r^2 = -1/2.$$

In fact, the same factor cancels from numerator and denominator as we take higher differences, if a sufficient number of zeros intervene. More explicitly stated, the formula (9) found for the data is valid also for $\Delta^k(n)^k$, provided $n \geqq t + 2$.

To make this more concrete, it may be noted that cycle lengths corresponding to $t = 0, 1, 2, 3$, and 4, are respectively

$$(20) \qquad 4, 3, 2.73, 2.60, 2.52.$$

From (15), we see directly that an element of $\Delta^k(n)^k$ is correlated only with certain other elements which are at distances from it which are multiples of n.

Some of the foregoing results may be included in a theorem as follows: THEOREM: *Given a sequence of independent chance variates, each subject to the normal distribution* (3) *with mean zero. Upon this material, let k summings or averagings by n be performed and k differencings, in any order. Then the resulting sequence has something of the same chaotic nature as the data. In particular for $n \geqq 2$ the expected frequency of changes of sign is the same,—viz.,* $1/4$ *for change from minus to plus and* $1/4$ *for change from plus to minus. Moreover, as n is increased from 2 to 3, 4, 5, $\cdots$, the expected frequency of other characteristics becomes the same, maxima and minima, points of inflection, etc., in accordance with* (9).

But, suppose now that after $k + 1$ summings by n, only k differencings are performed. Is the resulting sequence almost chaotic? Hardly so. At least, it can be shown that changes of sign in each direction have no longer an expected frequency fixed at $1/4$; but this expected frequency decreases as n increases. To show this, formula (5) is applied to (16); and setting in (10), $C = {}_{2k}C_k$, $C' = {}_{2k}C_{k-1}$ it follows that

$$(21) \qquad \cos \theta = [(n - 1)C - C']/nC = 1 - (2k + 1)/n(k + 1).$$

Then $\cos \theta > 1 - 2/n$; and the cycle length for expected changes of sign in definite direction is somewhat greater than that obtained by setting $\cos \theta = 1 - 2/n$. For values of n not too small, we may write $\cos \theta = 1 - \theta^2/2$, approximately; and then approximately

$$(22) \qquad \text{cycle length for definite change of sign in } \Delta^k(n)^{k+1} \text{ is } \pi\sqrt{n}.$$

If $n = 9$, this approximate length is 9.4, assuming k fairly large, whereas the more exact length is 9.2.

Consider now the result of summing $k + 2$ times, and then differencing only k

times. For this purpose, a few formulas for summing squares will be useful. By the method of differences it can be shown that if $l = a + nh$, and

(23) $T = a^2/2 + (a + h)^2 + (a + 2h)^2 + \cdots + (a + \overline{n-1}h)^2 + l^2/2,$

then

$$T = n(a^2 + al + l^2)/3 + (l - a)^2/6n. \tag{24}$$

Suppose, now, that a/n takes on the values $0, {}_kC_0, -{}_kC_1, \cdots, (-1)^k{}_kC_k$ in succession, while l/n takes on the values ${}_kC_0, -{}_kC_1, \cdots, (-1)^k{}_kC_k, 0$. Let U be the sum of the $(k + 1)$ values of T thus obtained. Then by (10).

$$U = n^3(2\,{}_{2k}C_k - {}_{2k}C_{k-1})/3 + n\sum_{i=0}^{k+1} {}_{k+1}C_i^2/6. \tag{25}$$

$$U = \frac{n^3}{3}\frac{(k+2)(2k)!}{k!(k+1)!} + \frac{n}{6}\,{}_{2k+2}C_{k+1}. \tag{26}$$

Now, by applying to (16) one more summation by n, there are formed $(k + 2)$ arithmetic progressions of $(n + 1)$ terms each, alternately increasing and decreasing. The maximum and minimum terms at the juncture of the progressions are to be split into two halves to apply (23). Then the sum of the squares of these coefficients is given by (26). This forms a denominator for (5).

To obtain the numerator for (5) we note that from $ab = [a^2 + b^2 - (a - b)^2]/2$ it follows that if

(27) $V = a(a + h) + (a + h)(a + 2h) + \cdots + (a + \overline{n-1}h)(a + nh);$

then, from (23),

$$V = T - nh^2/3 = T - (l - a)^2/3n. \tag{28}$$

If now W is the sum of such V's, reference to the last terms of (24) and (26) shows that

$$W = U - (n/3)_{2k+2}C_{k+1}. \tag{29}$$

And hence, from (5),

$$\cos\theta = \frac{(k+2)n^2 - 4k - 2}{(k+2)n^2 + 2k + 1}. \tag{30}$$

Then

$$\cos\theta > \frac{n^2 - 4}{n^2 + 2}; \tag{31}$$

but only slightly greater when k is large. Again

$$\cos\theta > 1 - 6/n^2; \tag{32}$$

but only slightly greater when n is not small. In this case, $\cos\theta = 1 - \theta^2/2$, approximately. And thus, approximately, for large k, and for n not small

(33) cycle length for definite change of sign of $\Delta^k(n)^{k+2} = 1.81n$.

This gives for $n = 10$ a cycle length of 18.1; whereas, if $\cos\theta$ is taken as the right member of (31), the cycle length is 18.2.

Thus, if a $(k + 2)$-fold summation or averaging of random data is followed

by only k differencings, the resulting graduation or linear processing $z = \Delta^k(n)^{k+2}$ is decidedly not as chaotic as the data; as seen from (31) and (33). But further, $\Delta z = \Delta^{k+1}(n)^{k+2}$; and thus from (22) the cycle length for the expected maxima of z is about $\pi\sqrt{n}$.

Now Slutzky [1, p. 109] distinguished conspicuous waves from inconsequential "ripples." On this basis, the frequency of significant cyclical features for a chance variable, such as z, would be less than the frequency of the maxima. It is not so clear that the frequency of significant features of a chance variable will be *greater* than that for changes of sign in definite direction. That turned out to be true for graduated values such as discussed in my earlier paper [3, p. 262]. If this be also valid for z, we would expect that conspicuous "waves" of $\Delta^k(n)^{k+2}$ would have average length between $\pi\sqrt{n}$ and $1.81n$, except for small values of n and k.

4. Graduations or linear processes and their successive differences. If double summation by n is followed by a single differencing, the result—as indicated in (14)—is, for $n = 3$,

$$y_i = -x_i - x_{i+1} - x_{i+2} + x_{i+3} + x_{i+4} + x_{i+5}. \tag{34}$$

Then

$$y_{i+3} = -x_{i+3} - x_{i+4} - x_{i+5} + x_{i+6} + x_{i+7} + x_{i+8}. \tag{35}$$

Thus y_i and y_{i+3} are negatively correlated; since x_{i+3}, x_{i+4}, and x_{i+5} appear in each, but with sign changed. This would seem to tend to make maxima alternate with minima at distances of about 3; or at distances of n, in the general case (14). Here, following Slutzky and Romanovsky, the coefficient of correlation r_p between elements at a distance of p is taken as

$$r_p = E(x_r \cdot x_{r+p})/E(x_r)^2. \tag{36}$$

Using computed averages, instead of expected values, Alter [6] recommends a "correlation periodogram," in which r_p is the ordinate for abscissa p.

Moreover, we would expect a graduation (4) with coefficients a_i proportional to the ordinates y of the sinusoid $y = \sin(\alpha + 2\pi x/p)$ taken for $x = 1, 2, 3, \cdots$ to impress upon random data oscillations with maxima separated from minima by about $p/2$. But such a_i, as well as those in (34), have abrupt endings which introduce noticeable alterations. More satisfactory results come from tapering ends, such as appear in damped vibration, with coefficients about proportional to $e^{-c|x|} \cos 2\pi x/p$ or to $e^{-c|x|} \sin 2\pi x/p$. H. Labrouste and Mrs. Labrouste [7] give a powerful operator of this description.

Slutzky (loc. cit. pp. 119–123), Yule [8], and Walker [9] make use of damped harmonic vibration to explain the creation of cycles; while Bartels [10] approaches by a different method the oscillations that do not last.

Now the common graduation formulas have coefficients not conforming strictly to damped vibration, as the tapering ends vibrate more quickly. However, these ends have little more than a smoothing or stabilizing effect. Furthermore,

the coefficients for first differences are likely to conform to something like $e^{-c|x|} \sin 2\pi x/p$. Some *experimental* evidence will be presented for the following conclusion:

If the coefficients a_i *of a graduation or linear process* (4) *appear to conform roughly to equidistant ordinates of a damped vibration,* $\pm e^{-c|x|} \cos 2\pi x/p$ *or* $\pm e^{-c|x|} \sin 2\pi x/p$, *with changes of sign at intervals of* $p/2$, *then when this process* (4) *is applied to independent chance data having zero mean and constant variance, there is a tendency for the graduated or processed values to change sign at intervals of about* $p/2$.

A number of standard graduations have first and second differences—see (6), (7)—which bear a decided resemblance to damped vibrations, while the third or fourth differences have only moderate, if any, cyclic appearance. This is especially true of those graduations which are constructed by applying three summings—the number of terms in a sum being in general different—and a fourth

TABLE I

Coefficients (×350) for Spencer 21-term graduation and for first four differences. Also theoretical cycle lengths for change in sign in values obtained from random data

			Cycle Length
Grad.	+	6,18,33,47,57,60,57,47,33,18,6	10.7
	−	1,3,5,5,2 2,5,5,3,1	
1st D.	+	1,2,2,0 3,10,14,15,12,8,3	7.0
	−	3,8,12,15,14,10,3 0,2,2,1	
2nd D.	+	2,3,5,4,3 3,4,5,3,2	5.5
	−	1,1,0 1,4,7,6,7,4,1 0,1,1	
3rd D.	+	1,0 1,1,4,3,3 1 2,1,2,1	3.2
	−	1,2,1,2, 1 3,3,4,1,1 0,1	
4th D.	+	1,1,1 1 0 1 4 4 1 0 1 1,1,1	1.6
	−	1 1 3 3 0 2 0 3 3 1 1	

process with negative coefficients. This is, indeed, a favorite form of graduation, with which are associated the names of Woolhouse, Spencer, Higham, Kenchington, Henderson, etc. The Spencer 21-term formula, for which some features have already been described, [3, p. 262], will now be examined, with special reference to its differences. Cycle length for change of sign is one-half that for change from minus to plus.

In the graduation formula, itself, there are 11 positive coefficients, centrally located, and relatively large as compared with the negative coefficients. This 11 is close to 10.7 the theoretical cycle length for changes of sign of $y_r - 4.5$, the difference between the graduated value y_r and its mean—the arithmetic mean of 1, 2, $\cdots$, 9. The structure of the first and second differences also

matches closely the corresponding cycle lengths. In the third differences, there is a break at the center; but still there appears considerable regularity. But among fourth differences, the zigzag is the prominent feature. Now the theorem of Section 3 does not really apply to the Spencer formula, with its two summations by fives and one summation by sevens, and another process. But it is not surprising that the cyclicity ceases after passing the third differences.

As a basis for comparing observed values with expected values, the tenth digits in the 600 logarithms from log 200 to log 799 were taken as a random set of numbers. These 600 numbers had been given a Spencer 21-term graduation [3, pp. 261–262], yielding 580 graduated values. From these the 579 first differences were found, the 578 second differences, etc. These numbers, 580, 579, $\cdots$, were multiplied respectively by the expected relative frequences of change in sign of $y_r - 4.5$, of Δy_r, $\Delta^2 y_r$, etc., as found by use of (5), (8), and similar expressions to form the following table.

The most abrupt change in frequency or cycle length appears to occur in passing from third to fourth differences. In Table I, this is seen in the configura-

TABLE II

Comparison of expected changes of sign with observed changes for a Spencer 21-term graduation

	Expected Number of Changes from − to +	*Observed Number of Changes from − to +*
Graduated values—4.5	27.2	27
First differences	41.3	42
Second differences	52.9	48
Third differences	90.4	74
Fourth differences	176.7	146

tion of positive and negative terms, and in the drop from 3.2 to 1.6 in cycle length; and in Table II in the corresponding increase in expected sign changes from 90.4 to 176.7. More spectacular is the increase in the number of zigzags represented by $-$, $+$, $-$, $+$. Among the third differences, there were found only 13 instances of four successive terms with signs as just indicated, whereas among fourth differences there were found 75 such instances. For random material, about 36 such zigzags would be expected—decidedly more than found among the third difference, and decidedly less than found among the fourth differences.

The Spencer 21-term graduation appears to be fairly representative of commonly used graduations as regards regularity or irregularity in the distribution of positive and negative coefficients among the differences. For graduations with a much larger number of terms, the alternation of sign in fourth differences may not be so rapid, as, e.g. in the 35-term 5th degree parabolic graduation which Macaulay [11] calls No. 18. On the other hand, for a formula with non-tapering ends, such as the 13-term formula which Macaulay gives [11,

p. 64], the coefficients appearing in the differences are more irregular, especially at the ends. While the Spencer formula is fairly representative, different formulas have distinguishing features. If it is desirable to form an idea of what a given formula will do to random data, a table like Table I can be constructed.

5. Summary. When upon independent chance data, summing, averaging or some more general graduation process is used, the graduated values tend to assume a wavy configuration. These waves often seem to have a fair amount of regularity or cyclicity. The first differences usually, and often other differences of the graduated values, are decidedly cyclic. But, as we go in turn to the higher differences, the cyclicity may weaken. Indeed there may be a return to something like randomness. And subsequent differencings may tend to set up zigzags.

If $(k + 2)$ successive summings by n have been performed on independent chance data, with n not too small, say $n \geqq 5$—then $k + 2$ differencings will just about bring back the original chaotic or random condition. But with only k or $(k + 1)$ differencings, a definite cyclicity remains, at least theoretically, in the expected values.

In the case of the Spencer 21-term graduation, the coefficients for the successive differences indicate the appearance of cyclicity in first, second, and third differences.

REFERENCES

[1] Eugen Slutzky, "The summation of random causes as the source of cyclic processes," *Econometrica*, Vol. 5 (1937), pp. 105-146. This supplements an earlier paper (1927) in Russian.

[2] V. Romanovsky, "Sur la loi sinusoïdale limite," *Rendiconti del Circolo Matematico di Palermo*, Vol. 56 (1932), pp. 82-111.

[3] Edward L. Dodd, "The length of the cycles which result from the graduation of chance elements," *Annals of Math. Stat.*, Vol. 10 (1939), pp. 254-264.

[4] Edward L. Dodd, "The problem of assigning a length to the cycle to be found in a simple moving average and in a double moving average of chance data," *Econometrica*, Vol. 9 (1941), pp. 25-37.

[5] Robert Henderson, *Graduation of Mortality and Other Tables*, Actuarial Society of America, New York, 1919.

[6] Dinsmore Alter, "A group or correlation periodogram, with applications to the rainfall of the British Isles," *Monthly Weather Review*, Vol. 55 (1927), pp. 263-266.

[7] H. and Mrs. Labrouste, "Harmonic analysis by means of linear combinations of ordinates," *Terr. Mag. and Atmos. Elec.*, Vol. 41 (1936), pp. 17-18.

[8] G. Udny Yule, "On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers," *Phil. Trans.* A, Vol. 226 (1927), pp. 267-298.

[9] Sir Gilbert Walker, "On periodicity in series of related terms," *Roy. Soc. Proc.*, Ser. A, Vol. 131 (1931), pp. 518-532.

[10] J. Bartels, "Random fluctuations, persistence, and quasi-persistence in geophysical and cosmical periodicities," *Terr. Mag. and Atmos. Elec.*, Vol. 40 (1935), pp. 1-60.

[11] F. R. Macaulay, *The Smoothing of Time Series*, Publication of the National Bureau of Economic Research, No. 19, New York, 1931.

ON THE DISTRIBUTION OF WILKS' STATISTIC FOR TESTING THE INDEPENDENCE OF SEVERAL GROUPS OF VARIATES

By A. Wald[1] and R. J. Brookner[1]

Columbia University

1. **Introduction.** We consider p variates $x_1, x_2, \cdots, x_p$ which have a joint normal distribution. Let the variates be divided into k groups; group one containing $x_1, x_2, \cdots, x_{p_1}$, group two containing $x_{p_1+1}, x_{p_1+2}, \cdots, x_{p_2}$, etc. We are interested in testing the hypothesis that the set of all population correlation coefficients between any two variates which belong to different groups is zero.

Wilks[2] has derived, by using the Neyman-Pearson likelihood ratio criterion, a statistic based on N independent observations on each variate with which one may test this hypothesis. Let $\| r_{ij} \|$ be the matrix of sample correlation coefficients; Wilks' statistic, λ, is the ratio of the determinant of the p-rowed matrix of sample correlations to the product of the p_1-rowed determinant of correlations of the variates of group one, the $(p_2 - p_1)$-rowed determinant of correlations of the second group, etc. That is

$$\lambda = \frac{| r_{ij} |}{| r_{\alpha_1\beta_1} | \cdot | r_{\alpha_2\beta_2} | \cdots | r_{\alpha_k\beta_k} |}$$

where $| r_{\alpha_i\beta_i} |$ is the principal minor of $| r_{ij} |$ corresponding to the ith group.

In order to use the test, the distribution function of λ must be known. Wilks has shown that in certain cases the exact distribution is a simple elementary function; in other cases it is an elementary function, but one which is rather unwieldy and which does not lend itself readily to practical use. It is our purpose in this paper (1) to show a method by which the exact distribution can be explicitly given as an elementary function for a certain class of groupings of the variates, and (2) to give an expansion of the exact cumulative distribution function in an infinite series which is applicable to any grouping.

2. **The exact distribution of λ.** By the method to be described, the exact distribution of λ can be found when the numbers of variates in the groups are such that there are an odd number in at most one group. If the number of variates is small, say at most eight, the method will increase only slightly the list of distribution functions that Wilks gives in his paper.

---

[1] Research under a grant-in-aid of the Carnegie Corporation of New York.

[2] S. S. Wilks, "On the independence of k sets of normally distributed statistical variables," *Econometrica*, Vol. 3 (1935), pp. 309-326. Other references to Wilks in this paper except where otherwise noted are to this publication.

For purposes of deriving the distribution of λ we may assume that $E(x_u) = 0$, $(u = 1, 2, \cdots, p)$; that there are $n = N - 1$ independent observations $x_{u\alpha}$ $(\alpha = 1, 2, \cdots, n)$ on each variate x_u; and that the sample covariance between x_i and x_j is given by $s_{ij} = \sum_{\alpha=1}^{n} x_{i\alpha}x_{j\alpha}/n$. We define u' (a function of u) to be the total number of variables in all the groups which precede the group in which x_u lies. The complete theory is independent of the ordering of the groups and of the ordering of the variates within the groups; hence without loss of generality, we may assume that if any group contains an odd number of variates, it will be the last group, hence u' is always an even integer.

Wilks has shown that λ is a product $\prod_{u=p_1+1}^{p} z_u$ where each z_u is distributed independently of the others, and that the distribution of z_u is

$$\frac{z_u^{\frac{1}{2}(n-u-1)}(1 - z_u)^{\frac{1}{2}(u'-2)}}{B[\frac{1}{2}(n - u + 1), u'/2]} dz_u . \tag{1}$$

Now let $y_u = \log z_u$, then the characteristic function of y_u is

$$\begin{aligned}\phi_u(t) &= \frac{1}{B[\frac{1}{2}(n - u + 1), u'/2]} \int_0^1 e^{t \log z_u} z_u^{\frac{1}{2}(n-u-1)}(1 - z_u)^{\frac{1}{2}(u'-2)} dz_u \\ &= \frac{1}{B[\frac{1}{2}(n - u + 1), u'/2]} \int_0^1 z_u^{\frac{1}{2}(n-u-1)+t}(1 - z_u)^{\frac{1}{2}(u'-2)} dz_u\end{aligned}$$

where t is a pure imaginary. It is known[3] that this integral, even with complex exponents, is the Beta-function so long as the real parts of both exponents are greater than minus one, so

$$\begin{aligned}\phi_u(t) &= \frac{B[\frac{1}{2}(n - u + 1) + t, u'/2]}{B[\frac{1}{2}(n - u + 1), u'/2]} \\ &= \frac{\Gamma[\frac{1}{2}(n - u + 1) + t\cdot]\Gamma[\frac{1}{2}(n - u + 1 + u')]}{\Gamma[\frac{1}{2}(n - u + 1 + u') + t]\cdot\Gamma[\frac{1}{2}(n - u + 1)]}.\end{aligned} \tag{2}$$

But here u' is always an even integer, hence by the well known recursion formula of the Gamma-function, which is valid for complex arguments excluding only negative integers

$$\phi_u(t) = c_u\{[\tfrac{1}{2}(n - u + 1) + t][\tfrac{1}{2}(n - u + 3) + t] \cdots [\tfrac{1}{2}(n - u + u' - 1) + t]\}^{-1}$$

where

$$c_u = [\tfrac{1}{2}(n - u + 1)][\tfrac{1}{2}(n - u + 3)] \cdots [\tfrac{1}{2}(n - u + u' - 1)].$$

[3] See Whittaker and Watson, *A Course in Modern Analysis*, Fourth edition 1927, Chap. 12.

Now set

$$y = \log \lambda = y_{p_1+1} + y_{p_1+2} + \cdots + y_p$$

and the characteristic function of y is

$$\phi(t) = \prod_{u=p_1+1}^{p} c_u \{[\tfrac{1}{2}(n - u + 1) + t][\tfrac{1}{2}(n - u + 3) + t] \cdots [\tfrac{1}{2}(n - u + u' - 1) + t]\}^{-1}.$$

From the characteristic function, we can obtain the distribution function, $g(y)$, of y by the relation

$$g(y) = \frac{c_n}{2\pi i} \int_{-i\infty}^{i\infty} \frac{e^{-yt}\, dt}{\prod_{u=p_1+1}^{p} [\tfrac{1}{2}(n - u + 1) + t] \cdots [\tfrac{1}{2}(n - u + u' - 1) + t]}$$
$$= \frac{c_n}{2\pi i} \int_{-i\infty}^{i\infty} \Phi(t)\, dt,$$

where

$$c_n = \prod_{u=p_1+1}^{p} c_u.$$

The integration can be carried out by the method of residues; since y is always negative (the range of λ is from 0 to 1), on a half circle with center at the origin in the negative half of the complex t-plane, the integral of the function $\Phi(t)$ converges to zero as the radius of the circle becomes infinite. Since $\Phi(t)$ is analytic except for a finite number of poles on the negative real axis, $g(y)$ is c_n times the sum of the residues at these points.

Now $\Phi(t)$ is of the form $\dfrac{e^{-yt}}{P(t)}$ where $P(t)$ is a polynomial in t as follows: suppose that the groups contain $r_1, r_2, \cdots, r_k$ variables respectively, then let $(k_j + 1)$ be the number of these r's which are greater than or equal to j; then

$$P(t) = [\tfrac{1}{2}(n - 2) + t]^{k_1}[\tfrac{1}{2}(n - 3) + t]^{k_2}[\tfrac{1}{2}(n - 4) + t]^{k_3+k_1}[\tfrac{1}{2}(n - 5) + t]^{k_4+k_2}$$
$$[\tfrac{1}{2}(n - 6) + t]^{k_5+k_3+k_1} \cdots [\tfrac{1}{2}(n - p + 1) + t]^{k_{p-2}+k_{p-4}+\cdots+k_{[\frac{1}{2}p]-[\frac{1}{2}(p-3)]}}.$$

where

$$[\sigma/2] = \begin{array}{l} \sigma/2 \text{ if } \sigma \text{ is even} \\ (\sigma - 1)/2 \text{ if } \sigma \text{ is odd.} \end{array}$$

Then

$$g(y; r_1, r_2, \cdots, r_k) = c_n \sum_{\alpha=1}^{p-2} \frac{1}{\theta_\alpha!} \frac{d^{\theta_\alpha}}{dt^{\theta_\alpha}} [(t + \tfrac{1}{2}(n - \alpha - 1))^{\theta_\alpha+1} \Phi(t)]_{t=-\frac{1}{2}(n-\alpha-1)}$$

where

$$\theta_\alpha + 1 = k_\alpha + k_{\alpha-2} + \cdots + k_{[\frac{1}{2}(\alpha+2)]-[\frac{1}{2}(\alpha-1)]}.$$

It can be shown that θ_α is $\geqq 0$ for α between 1 and $p - 2$. Thus we have $g(y; r_1, r_2, \cdots, r_k)$ and from it we can calculate $f(\lambda; r_1, r_2, \cdots, r_k)$.

Suppose $p = 8$ and that the variables are divided into two groups of four each, then we will calculate the distribution function $f(\lambda; 4, 4)$. Now

$$g(y; 4, 4) = \frac{c_n}{2\pi i} \int_{-i\infty}^{i\infty} \frac{e^{-yt}\, dt}{[\frac{1}{2}(n-2) + t][\frac{1}{2}(n-3) + t][\frac{1}{2}(n-4) + t]^2 \cdot [\frac{1}{2}(n-5) + t]^2[\frac{1}{2}(n-6) + t][\frac{1}{2}(n-7) + t]}$$

and

$$c_n = \left(\frac{n-2}{2}\right)\left(\frac{n-3}{2}\right)\left(\frac{n-4}{2}\right)^2\left(\frac{n-5}{2}\right)^2\left(\frac{n-6}{2}\right)\left(\frac{n-7}{2}\right).$$

Then

$$g(y; 4, 4) = 16c_n\left[\frac{-e^{\frac{1}{2}(n-2)y}}{90} + e^{\frac{1}{2}(n-3)y} + \frac{8e^{\frac{1}{2}(n-4)y}}{9} - \frac{8e^{\frac{1}{2}(n-5)y}}{9} - e^{\frac{1}{2}(n-6)y} + \frac{e^{\frac{1}{2}(n-7)y}}{90} - \frac{ye^{\frac{1}{2}(n-4)y}}{3} + \frac{ye^{\frac{1}{2}(n-5)y}}{3}\right].$$

Since

$$y = \log \lambda, \qquad dy = \frac{d\lambda}{\lambda},$$

we have

$$f(\lambda; 4, 4,) = \frac{16c_n}{3}\left[-\frac{\lambda^{\frac{1}{2}(n-4)}}{30} + \frac{\lambda^{\frac{1}{2}(n-5)}}{2} - \frac{8\lambda^{\frac{1}{2}(n-6)}}{3} + \frac{8\lambda^{\frac{1}{2}(n-7)}}{3} - \frac{\lambda^{\frac{1}{2}(n-8)}}{2} + \frac{\lambda^{\frac{1}{2}(n-9)}}{30} - (\lambda^{\frac{1}{2}(n-7)} + \lambda^{\frac{1}{2}(n-6)}) \log \lambda\right].$$

The cumulative distribution function is given by

$$\begin{aligned} J_w(4, 4) &= \text{Prob } [\lambda \leq w; 4, 4] \\ &= \frac{16c_n}{3} w^{\frac{1}{2}(n-7)}\left[\frac{1}{15(n-7)} - \frac{w^{\frac{1}{2}}}{n-6} - \frac{4(4n-23)w}{3(n-5)^2} + \frac{14(4n-13)w^{\frac{3}{2}}}{3(n-4)^2} + \frac{w^2}{n-3} - \frac{w^{\frac{5}{2}}}{15(n-2)} - \left(\frac{2w}{n-5} + \frac{2w^{\frac{3}{2}}}{n-4}\right)\log w\right]. \end{aligned}$$

Wilks' expression for the cumulative distribution function appears to be quite different, but if we substitute $n = N - 1$ and use the relation

$$\begin{aligned} \beta_{\sqrt{w}}(N-6; 4) &= \frac{\Gamma(N-2)}{\Gamma(N-6)\cdot\Gamma(4)} \int_0^{\sqrt{w}} x^{N-7}(1-x)^3\, dx \\ &= \tfrac{1}{6}(n-2)(n-3)(n-4)(n-5) \cdot \left[\frac{w^{\frac{1}{2}(n-5)}}{n-5} - \frac{3w^{\frac{1}{2}(n-4)}}{n-4} + \frac{3w^{\frac{1}{2}(n-3)}}{n-3} - \frac{w^{\frac{1}{2}(n-2)}}{n-2}\right] \end{aligned}$$

it can be shown that the two formulas for the cumulative distribution are identical.

In cases where u' is not always an even integer, the exact distribution function of λ can still be obtained using this method. However, in such a case, the gamma functions do not cancel out and the integrand has an infinitude of poles, so the function is expressed by an infinite series. We will use a different method to obtain an infinite series expansion.

3. **A series expansion of the cumulative distribution function.** Let us put $v = -y$, and let the density function of v be $h(v)$, then from (2), we have

$$h(v)\,dv = dv\,\frac{c_n}{2\pi i}\int_{-i\infty}^{i\infty} e^{vt} \prod_{u=r_1+1}^{p} \frac{\Gamma[\frac{1}{2}(n-u+1)+t]\,dt}{\Gamma[\frac{1}{2}(n-u+1+u')+t]}.$$

Since v is a monotonic decreasing function of λ, and since the critical region for testing the null hypothesis is given by the inequality $\lambda < \lambda_0$, then the critical region will be defined by $v > v_0$, where v_0 is such that

$$\int_{v_0}^{\infty} h(v)\,dv$$

is equal to a chosen level of significance.

PROPOSITION 1.

$$h(v) = h_n(v)\bar{\psi}(v)$$

where $\bar{\psi}(v)$ *does not depend on* n, *and* $h_n(v) = c_n e^{-\frac{p}{2}v}$.

PROOF: Let

$$t' = t + \tfrac{1}{2}(n-p).$$

Then

$$h(v) = \frac{c_n}{2\pi i}\int_{-i\infty+\frac{1}{2}(n-p)}^{i\infty+\frac{1}{2}(n-p)} e^{v(t'-\frac{1}{2}(n-p))} \prod_{u} \frac{\Gamma[\frac{1}{2}(p-u+1)+t']\,dt'}{\Gamma[\frac{1}{2}(p-u+u'+1)+t']}.$$

Now the area in the complex plane bounded by the vertical line through $\frac{1}{2}(n-p)$, by the vertical line through the origin, and by arcs of a circle with center at the origin of arbitrary radius is one in which the integrand is everywhere regular. Furthermore, the integral along the arcs approaches zero as the radius of the circle approaches infinity, hence the integrals along the vertical line through $\frac{1}{2}(n-p)$ and along the vertical axis are equal. Then we may write

$$\begin{aligned}\frac{e^{\frac{p}{2}v}}{c_n}h(v) &= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{v(t'+p/2)} \prod_{u} \frac{\Gamma[\frac{1}{2}(p-u+1)+t']\,dt'}{\Gamma[\frac{1}{2}(p-u+u'+1)+t']}\\ &= \bar{\psi}(v).\end{aligned}$$

Therefore

$$h(v) = c_n e^{-\frac{p}{2}v}\bar{\psi}(v).$$

PROPOSITION 2.

$$I = \lim_{n\to\infty} \int_0^\infty \frac{c_n e^{-\frac{n}{2}v} v^{r-1}\, dv}{\Gamma(r)} = 1$$

where we define

$$r = \sum_{j=i+1}^{k} \sum_{i=1}^{k-1} \frac{r_i r_j}{2}$$

so that

$$\begin{aligned} r &= \tfrac{1}{2}[r_2 r_1 + r_3(r_1 + r_2) + \cdots + r_k(r_1 + r_2 + \cdots + r_{k-1})] \\ &= \tfrac{1}{2} \sum_u u'. \end{aligned}$$

PROOF: Let

$$\frac{n}{2} v = v^*$$

then

$$\begin{aligned} \int_0^\infty c_n e^{-\frac{n}{2}v} v^{r-1}\, dv &= \int_0^\infty c_n e^{-v^*} \left(\frac{2}{n}\right)^r (v^*)^{r-1}\, dv^* \\ &= c_n \left(\frac{2}{n}\right)^r \Gamma(r). \end{aligned}$$

Hence

$$I = \lim_{n\to\infty} c_n \left(\frac{2}{n}\right)^r$$

but

$$c_n = \prod_u \frac{\Gamma\frac{1}{2}(n - u + 1 + u')}{\Gamma\frac{1}{2}(n - u + 1)}$$

and therefore

$$I_u = \lim_{n\to\infty} \frac{\Gamma\frac{1}{2}(n - u + 1 + u')}{\Gamma\frac{1}{2}(n - u + 1)} \left(\frac{2}{n}\right)^{u'/2} = 1$$

by an application of the Stirling approximation. Therefore

$$I = \prod_u I_u = 1.$$

We then write

$$\psi(v) = \frac{\bar{\psi}(v)\Gamma(r)}{v^{r-1}}$$

hence

(3) $$h(v) = \frac{c_n e^{-\frac{1}{2}v} v^{r-1} \psi(v)}{\Gamma(r)}.$$

PROPOSITION 3. *For any positive integer s,*

$$\lim_{n\to\infty} \left\{ n^s \cdot \text{Prob}\left(v > \frac{1}{\sqrt{n}}\right)\right\} = 0.$$

PROOF: Since $v = -\log \lambda$, the inequality $v > 1/\sqrt{n}$ is equivalent to the inequality $\lambda < e^{-1/\sqrt{n}}$. Since $\lambda = \prod_{u=p_1+1}^{p} z_u$, the inequality $\lambda < e^{-1/\sqrt{n}}$ implies that there exists at least one value of u for which

$$z_u < e^{-1/(p-p_1)\sqrt{n}}.$$

Hence

$$\sum_{u=p_1+1}^{p} P(z_u < e^{-1/(p-p_1)\sqrt{n}}) \geq P(\lambda < e^{-1/\sqrt{n}}) = P(v > 1/\sqrt{n}).$$

Hence in order to prove Proposition 3 we have only to show that for each u and any arbitrary positive integer s

$$\lim_{n\to\infty} \{n^s \cdot P(z_u < e^{-1/(p-p_1)\sqrt{n}})\} = 0.$$

From (1) we have

$$P(z_u < e^{-1/(p-p_1)\sqrt{n}})$$

$$= \frac{1}{B[\frac{1}{2}(n-u+1); u'/2]} \int_0^{e^{-1/(p-p_1)\sqrt{n}}} z_u^{\frac{1}{2}(n-u-1)} (1-z_u)^{\frac{1}{2}(u'-2)}\, dz_u.$$

Over the range of integration, we have $z_u \leq e^{-1/(p-p_1)\sqrt{n}}$ so

$$P(z_u < e^{-1/(p-p_1)\sqrt{n}}) \leq \frac{e^{\frac{1}{2}(n-u-1)/(p-p_1)\sqrt{n}}}{B[\frac{1}{2}(n-u+1); u'/2]} \int_0^{e^{-1/(p-p_1)\sqrt{n}}} (1-z_u)^{\frac{1}{2}(u'-2)}\, dz_u$$

$$= \frac{e^{-\frac{1}{2}(n-u-1)/(p-p_1)\sqrt{n}}}{B[\frac{1}{2}(n-u+1); u'/2]} \left[-\frac{2}{u'} (1-z_u)^{u'/2} \right]_0^{e^{-1/(p-p_1)\sqrt{n}}}$$

$$= \frac{2e^{-\frac{1}{2}(n-u-1)/(p-p_1)\sqrt{n}}}{u' \cdot B[\frac{1}{2}(n-u+1); u'/2]} [1 - (1 - e^{-1/(p-p_1)\sqrt{n}})^{u'/2}].$$

It follows from the Stirling formula that

$$\lim_{n\to\infty} \left(\frac{n}{2}\right)^{u'/2} B[\tfrac{1}{2}(n-u+1); u'/2] = \lim_{n\to\infty} \frac{\Gamma\frac{1}{2}(n-u+1)\Gamma(u'/2)}{\Gamma\frac{1}{2}(n-u+u'+1)} \left(\frac{n}{2}\right)^{u'/2}$$
$$= \Gamma(u'/2).$$

Since

$$\lim_{n \to \infty} n^{\frac{1}{2}u'+s} e^{-\sqrt{n}/2(p-p_1)} = 0$$

and

$$\lim_{n \to \infty} (1 - (1 - e^{-1/\sqrt{n}})) = 1,$$

the proposition follows.

PROPOSITION 4. *The function $\psi(v)$ of formula* (3) *can be expanded in a power series, i.e.*

$$\psi(v) = \alpha_0 + \alpha_1 v + \alpha_2 v^2 + \cdots$$

with a finite radius of convergence.

PROOF: Wilks[4] has considered the following integral equation:

$$\int_0^B w^t g(w)\, dw = CB^t \frac{\Gamma(b_1 + t) \cdot \Gamma(b_2 + t) \cdots \Gamma(b_q + t)}{\Gamma(c_1 + t) \cdot \Gamma(c_2 + t) \cdots \Gamma(c_q + t)},$$

where $C = \dfrac{\Gamma(c_1) \cdot \Gamma(c_2) \cdots \Gamma(c_q)}{\Gamma(b_1) \cdot \Gamma(b_2) \cdots \Gamma(b_q)}$, B and $g(w)$ are independent of t, and $b_i < c_i$ $(i = 1, 2, \cdots, q)$. Wilks has shown that the solution of the integral equation, $g(w)$, is given by the following expression:

$$\begin{aligned} g(w) = \frac{kw^{b_q-1}\left(1 - \frac{w}{B}\right)^{\gamma_q - \beta_q - 1}}{B^{b_q}} & \int_0^1 \int_0^1 \cdots \int_0^1 v_1^{c_1-b_1-1} v_2^{c_2-b_2-1} \cdots v_{q-1}^{c_{q-1}-b_{q-1}-1} \\ & \times (1 - v_1)^{\gamma_{q-1}-\beta_{q-1}-1} (1 - v_2)^{\gamma_{q-2}-\beta_{q-2}-1} \cdots (1 - v_{q-1})^{\gamma_1-\beta_1-1} \\ & \times \left[1 - v_1\left(1 - \frac{w}{B}\right)\right]^{b_1-c_2} \left[1 - \{v_1 + v_2(1 - v_1)\}\left(1 - \frac{w}{B}\right)\right]^{b_2-c_3} \cdots \\ & \times \Big[1 - \{v_1 + v_2(1 - v_1) + \cdots \\ & \qquad + v_{q-1}(1 - v_1)(1 - v_2) \cdots (1 - v_{q-2})\}\left(1 - \frac{w}{B}\right)\Big]^{b_{q-1}-c_q} \\ & \times dv_1\, dv_2 \cdots dv_{q-1} \end{aligned} \tag{4}$$

where

$$k = \prod_{i=1}^{q} \frac{\Gamma(c_i)}{\Gamma(b_i)\Gamma(c_i - b_i)}$$

and

$$\gamma_i = \sum_{j=0}^{i-1} c_{q-j} \qquad \beta_i = \sum_{j=0}^{i-1} b_{q-j}$$

[4] S. S. Wilks, "Certain generalizations in the analysis of variance," *Biometrika*, Vol. 24 (1932), pp. 474–5.

the range of w being $0 \leqq w \leqq B$. Wilks has furthermore shown that

$$(5) \quad \{v_1 + v_2(1 - v_1) + \cdots + v_i(1 - v_1)(1 - v_2) \cdots (1 - v_{i-1})\}\left(1 - \frac{w}{B}\right) < 1$$

for $w > 0$ and $0 \leqq v_i \leqq 1$ $(i = 1, 2, \cdots, q - 1)$.

We denote the left hand side of (5) by ζ_i. The factor $(1 - \zeta_i)^{b_i - c_{i+1}}$ can be expanded in a power series, i.e.

$$(6) \quad (1 - \zeta_i)^{b_i - c_{i+1}} = (1 - \zeta_i)^{-(c_{i+1} - b_i)}$$
$$= 1 + (c_{i+1} - b_i)\zeta_i + \tfrac{1}{2}(c_{i+1} - b_i)(c_{i+1} - b_i + 1)\zeta_i^2 + \cdots$$

with a radius of convergence equal to one. Since we will show shortly that for the choices we make for the b_i's and c_i's, $c_{i+1} \geq b_i$, then all coefficients in this last expansion are non-negative. Substituting this series expansion (6) in (4), and ordering it according to powers of $(1 - w/B)$, the expression under the integral sign (in 4) becomes

$$(7) \quad \begin{aligned} &\theta_0(v_1, v_2, \cdots v_{q-1}) \\ &\quad + \theta_1(v_1, \cdots v_{q-1})\left(1 - \frac{w}{B}\right) + \theta_2(v_1, \cdots, v_{q-1})\left(1 - \frac{w}{B}\right)^2 + \cdots. \end{aligned}$$

This series is uniformly convergent over the domain defined by the inequalities $0 \leqq v_i \leqq 1$ $(i = 1, 2, \cdots, q - 1)$ and $|1 - w/B| < 1$. We can even say that (7) is uniformly convergent for $|1 - w/B| < 1$ if we substitute for each θ_i the maximum of θ_i with respect to $v_1, v_2, \cdots, v_{q-1}$. Hence we may integrate the series (7) with respect to $v_1, v_2, \cdots v_{q-1}$ term by term, i.e.

$$(8) \quad \int_0^1 \int_0^1 \cdots \int_0^1 (7)\, dv_1\, dv_2 \cdots dv_{q-1} = \sigma_0 + \sigma_1\left(1 - \frac{w}{B}\right) + \sigma_2\left(1 - \frac{w}{B}\right)^2 + \cdots$$

and the series (8) is uniformly convergent for $|1 - w/B| < 1$. The coefficients $\sigma_0, \sigma_1, \cdots$ are non-negative.

The case of the λ statistic which we are considering is a special case of this integral equation which we obtain by making the following substitutions:

$$w = \lambda, \quad B = 1, \quad u = r + p_1, \quad q = p - p_1$$
$$b_r = \tfrac{1}{2}(n - u + 1), \quad c_r = \tfrac{1}{2}(n - u + u' + 1), \quad (r = 1, 2, \cdots, p - p_1)$$

Note that then

$$c_{r+1} - b_r = \tfrac{1}{2}[(u + 1)' - 1] \geqq 0.$$

Hence, according to (4)

$$g(\lambda)\, d\lambda = k \cdot \lambda^{\frac{1}{2}(n-p-1)}(1 - \lambda)^{\frac{1}{2}\Sigma u' - 1}\{\sigma_0 + \sigma_1(1 - \lambda) + \sigma_2(1 - \lambda^2) + \cdots\}\, d\lambda$$

where the infinite series converges for $|1 - \lambda| < 1$.

Now $v = -\log \lambda$, or $\lambda = e^{-v}$, hence

$$h(v)\, dv = k \cdot e^{-\frac{1}{2}(n-p+1)v} v^{r-1}\left(\frac{1 - e^{-v}}{v}\right)^{r-1} \{\epsilon_0 + \epsilon_1 v + \epsilon_2 v^2 + \cdots\}\, dv$$

where the series $\{\epsilon_0 + \epsilon_1 v + \epsilon_2 v^2 + \cdots\}$ is obtained from the series $\{\sigma_0 + \sigma_1(1 - \lambda) + \cdots\}$ by substituting for $(1 - \lambda)$ the Taylor expansion of $(1 - e^{-v})$. The series $\{\epsilon_0 + \epsilon_1 v + \epsilon_2 v^2 + \cdots\}$ has a finite radius of convergence.[5]

Hence the function $\psi(v)$ can be written as

$$\psi(v) = A \cdot e^{\frac{1}{2}(p-1)v} \left(\frac{1 - e^{-v}}{v}\right)^{r-1} \{\epsilon_0 + \epsilon_1 v + \epsilon_2 v^2 + \cdots\}$$

where A denotes a constant factor. Then since $e^{\frac{1}{2}(p-1)v}\left(\frac{1 - e^{-v}}{v}\right)^{r-1}$ can be expanded in a Taylor series around $v = 0$, Proposition 4 is proved.

4. **Evaluation of the coefficients in the expansion of $\psi(v)$.** Let the series expansion of $\psi(v)$ be

$$\psi(v) = \alpha_0 + \alpha_1 v + \alpha_2 v^2 + \cdots$$

Then we have

$$\int_0^\infty \frac{c_n e^{-\frac{1}{2}nv} v^{r-1}}{\Gamma(r)} (\alpha_0 + \alpha_1 v + \alpha_2 v^2 + \cdots)\, dv \equiv 1.$$

Now let $v^* = \frac{n}{2} v$, then

$$\int_0^\infty \left(\frac{2}{n}\right)^r \frac{c_n e^{-v^*} v^{*r-1}}{\Gamma(r)} \left(\alpha_0 + \frac{2\alpha_1 v^*}{n} + \frac{4\alpha_2 v^{*2}}{n^2} + \cdots\right) dv^* \equiv 1.$$

Suppose that the asymptotic expansion of $\left(\frac{n}{2}\right)^r \frac{1}{c_n}$ is given by

$$\beta_0 + \frac{\beta_1}{n} + \frac{\beta_2}{n^2} + \cdots.$$

On account of Proposition 3, we have that the asymptotic expansion in powers of $1/n$ of

$$\int_0^{\sqrt{n}} \frac{e^{-v^*} v^{*r-1}}{\Gamma(r)} \left(\alpha_0 + \frac{2\alpha_1}{n} v^* + \frac{4\alpha_2}{n^2} v^{*2} + \cdots\right) dv^* \tag{9}$$

must be equal to the asymptotic expansion of $\left(\frac{n}{2}\right)^r \frac{1}{c_n}$. Since we may integrate in (9) term by term for sufficiently large n, we easily obtain

$$\alpha_0 = \beta_0, \qquad \alpha_1 = \frac{\beta_1}{2r}, \cdots \alpha_k = \frac{\beta_k}{2^k \cdot r(r+1) \cdots (r+k-1)}.$$

[5] See A. Gutzmer, *Theorie der Eindeutigen Analytischen Funktionen*, 1906, pp. 91–2.

The asymptotic expansion of $\left(\frac{n}{2}\right)^r \frac{1}{c_n}$ can be calculated in the following manner:

$$\left(\frac{n+2}{n}\right)^r \frac{c_n}{c_{n+2}} = \frac{\beta_0 + \frac{\beta_1}{n+2} + \frac{\beta_2}{(n+2)^2} + \cdots}{\beta_0 + \frac{\beta_1}{n} + \frac{\beta_2}{n^2} + \cdots}$$

and

$$\left(\frac{n+2}{n}\right)^r \frac{c_n}{c_{n+2}} = (1 + 2/n)^r \prod_u \frac{n-u+1}{n-u+u'+1}.$$

Equating the right hand members of these last two equations, and taking logs, we obtain

$$\log\left[\beta_0 + \frac{\beta_1}{n+2} + \frac{\beta_2}{(n+2)^2} + \cdots\right] = r \log(1 + 2/n) + \sum_u \log\left(1 - \frac{u-1}{n}\right)$$
$$- \sum_u \log\left(1 - \frac{u-u'-1}{n}\right) + \log\left(\beta_0 + \frac{\beta_1}{n} + \frac{\beta_2}{n^2} + \cdots\right).$$

Then we expand each term in a series of powers of $1/n$ and equate coefficients of $1/n^i$ for each i. We obtain the following formulae for the first five β's:

$$\beta_0 = 1$$

$$\beta_1 = r + \tfrac{1}{4}\sum_u (u-1)^2 - \tfrac{1}{4}\sum_u (u-u'-1)^2$$

$$\beta_2 = \beta_1 + \frac{\beta_1^2}{2} - \frac{2r}{3} + \frac{1}{12}\sum_u (u-1)^3 - \frac{1}{12}\sum_u (u-u'-1)^3$$

$$\beta_3 = -\tfrac{4}{3}\beta_1 - \beta_1^2 - \tfrac{1}{3}\beta_1^3 + \beta_1\beta_2 + 2\beta_2 + \tfrac{2}{3}r$$
$$+ \tfrac{1}{24}\sum_u (u-1)^4 - \tfrac{1}{24}\sum_u (u-u'-1)^4$$

$$\beta_4 = 2\beta_1 + 2\beta_1^2 + \beta_1^3 + \frac{\beta_1^4}{4} - 3\beta_1\beta_2 + \beta_1\beta_3 - \beta_1^2\beta_2 - 4\beta_2$$
$$+ \frac{\beta_2^2}{2} + 3\beta_3 - \frac{4}{5}r + \frac{1}{40}\sum_u (u-1)^5 - \frac{1}{40}\sum_u (u-u'-1)^5.$$

5. **Practical use of the series.** In practical applications, the value of the statistic, say λ_0, is calculated, and it is desired that we determine whether or not this value of the statistic falls into the critical region. That is, for a particular grouping of the variates, for a particular number of degrees of freedom, and for a chosen level of significance α, there is determined from the distribution of λ, a value λ^* such that

$$\text{Prob}\,[\lambda < \lambda^*] = \alpha,$$

and if $\lambda_0 < \lambda^*$ we reject the hypothesis that in the population from which the sample is taken all the correlation coefficients between variates in different groups are zero.

Since v is a monotonic decreasing function of λ we make the test by computing $v_0 = -\log \lambda_0$ and we reject the hypothesis if $v_0 > v^*$ where $v^* = -\log \lambda^*$. But this is equivalent to computing Prob $[v > v_0]$ and if this value is less than α we reject the hypothesis. Now

$$\begin{aligned}\text{Prob } [v > v_0] &= J_{v_0}(r_1, r_2, \cdots, r_k) \\ &= \frac{C_n}{\Gamma(r)} \int_{v_0}^{\infty} e^{-\frac{1}{2}nv} v^{r-1}(1 + \alpha_1 v + \alpha_2 v^2 + \cdots)\, dv.\end{aligned}$$

Setting $\frac{nv}{2} = z$

$$\text{Prob } [v > v_0] = \left(\frac{2}{n}\right)^r \frac{C_n}{\Gamma(r)} \int_{nv_0/2}^{\infty} e^{-z} z^{r-1}\left[1 + \alpha_1 \frac{2z}{n} + \alpha_2 \left(\frac{2}{n}\right)^2 z^2 + \cdots\right] dz.$$

On account of Proposition 3 we obtain an asymptotic expansion of Prob $[v > v_0]$ by integrating the right hand member of the above equation term by term. This can be expressed by means of the incomplete gamma function, which is tabulated[6] in the form

$$I(u, p) = \frac{\int_0^{u\sqrt{p+1}} v^p e^{-v}\, dv}{\Gamma(p + 1)}.$$

We obtain

$$\begin{aligned}\text{Prob } [v > v_0] = \left(\frac{2}{n}\right)^r c_n \Bigg\{&\left[1 - I\left(\frac{nv_0}{2\sqrt{r}}, r - 1\right)\right] \\ &+ \frac{\beta_1}{n}\left[1 - I\left(\frac{nv_0}{2\sqrt{r+1}}, r\right)\right] + \frac{\beta_2}{n^2}\left[1 - I\left(\frac{nv_0}{2\sqrt{r+2}}, r + 1\right)\right] + \cdots\Bigg\}.\end{aligned}$$

The values of the constant $K = \left(\frac{2}{n}\right)^r c_n$ and the values of $\beta_1, \beta_2, \beta_3, \beta_4$ are herein tabulated for any grouping which might be made on six or fewer variates. Some cases, such as groupings $(1, p - 1)$, in which case the distribution of λ is the distribution of the multiple correlation coefficient; and as the groupings $(2, p - 2)$, the exact distribution for which was given by Wilks as an incomplete Beta-function, are superfluous here. These cases are included only for the sake of completeness.

[6] K. Pearson (Editor), *Tables of the Incomplete Gamma Function*, Biometric Laboratory, London, 1922.

Table of the First Four β's

Grouping	r	β_1	β_2	β_3	β_4
2,1	1	2	4	8	16
1,1,1	1.5	2.75	6.28125	13.38281	27.57568
3,1	1.5	3.75	12.03125	36.91406	111.55225
2,2	2	5	19	65	211
2,1,1	2.5	5.75	23.53125	83.97656	279.50538
1,1,1,1	3	6.5	28.625	106.9375	366.39844
4,1	2	6	28	120	496
3,2	3	9	55	285	1351
3,1,1	3.5	9.75	62.53125	334.10156	1615.91163
2,2,1	4	11	77	439	2229
2,1,1,1	4.5	11.75	86.03125	506.16406	2628.23974
1,1,1,1,1	5	12.5	95.625	580.6875	3085.52344
5,1	2.5	8.75	55.78125	315.82031	1690.65282
4,2	4	14	125	910	5901
3,3	4.5	15.75	154.03125	1205.03906	8277.55226
4,1,1	4.5	14.75	136.28125	1015.50781	6693.45068
3,2,1	5.5	17.75	189.53125	1584.10156	11445.75538
2,2,2	6	19	214	1866	13947
3,1,1,1	6	18.5	203.625	1740.9375	12797.27344
2,2,1,1	6.5	19.75	229.03125	2042.16406	15530.08351
2,1,1,1,1	7	20.5	244.625	2230.1875	17257.64836
1,1,1,1,1,1	7.5	21.25	260.78125	2430.49219	19139.02892

Tables of the Constant $K = \left(\frac{2}{n}\right)^r C_n$

n	21	111	31	22	211	1111	41	311
10	.800	.738	.646	.560	.517	.477	.480	.310
11	.818	.761	.676	.595	.553	.515	.521	.352
12	.833	.780	.702	.625	.585	.548	.556	.390
13	.846	.796	.724	.651	.612	.576	.586	.424
14	.857	.810	.743	.674	.637	.602	.612	.455
15	.867	.822	.759	.693	.658	.624	.636	.482
16	.875	.833	.774	.711	.677	.645	.656	.508
17	.882	.843	.787	.727	.694	.663	.675	.531
18	.889	.851	.798	.741	.709	.679	.691	.552
19	.895	.859	.808	.754	.723	.694	.706	.571
20	.900	.866	.818	.765	.736	.708	.720	.589
22	.909	.878	.834	.785	.758	.732	.744	.620
24	.917	.888	.847	.802	.777	.752	.764	.647
26	.923	.896	.859	.817	.793	.770	.781	.671
28	.929	.903	.869	.829	.807	.785	.796	.691
30	.933	.910	.877	.840	.819	.798	.809	.710
35	.943	.922	.894	.862	.843	.825	.835	.747
40	.950	.932	.908	.879	.862	.846	.855	.776
45	.956	.940	.918	.892	.877	.862	.871	.799
50	.960	.946	.926	.902	.889	.875	.883	.818
55	.964	.950	.932	.911	.899	.886	.894	.833
60	.967	.954	.938	.918	.907	.895	.902	.846
65	.969	.958	.943	.924	.914	.903	.910	.858
70	.971	.961	.947	.930	.920	.910	.916	.867
80	.975	.966	.953	.938	.930	.921	.926	.883
90	.978	.970	.959	.945	.937	.929	.934	.896
100	.980	.973	.963	.951	.943	.936	.941	.906

Tables of the Constant K (ii)

n	221	2111	32	11111	51	42	33
10	.269	.248	.336	.229	.323	.168	.136
11	.310	.288	.379	.268	.369	.206	.171
12	.347	.325	.417	.304	.410	.243	.205
13	.381	.359	.451	.338	.445	.277	.237
14	.412	.390	.481	.368	.478	.309	.268
15	.441	.418	.508	.397	.506	.339	.297
16	.467	.444	.533	.423	.532	.367	.324
17	.490	.468	.556	.447	.555	.392	.350
18	.512	.490	.576	.470	.576	.416	.374
19	.532	.511	.595	.490	.596	.438	.396
20	.551	.530	.612	.510	.613	.459	.417
22	.584	.564	.642	.544	.644	.496	.455
24	.613	.593	.668	.575	.671	.529	.489
26	.638	.619	.691	.601	.694	.558	.519
28	.660	.642	.711	.625	.714	.584	.546
30	.680	.662	.728	.646	.731	.607	.570
35	.720	.704	.764	.689	.767	.654	.621
40	.751	.737	.791	.723	.794	.692	.661
45	.776	.763	.813	.751	.816	.722	.694
50	.797	.785	.830	.773	.833	.747	.721
55	.814	.803	.845	.792	.848	.768	.743
60	.828	.818	.857	.808	.860	.786	.762
65	.841	.831	.868	.822	.870	.801	.779
70	.852	.842	.877	.833	.879	.814	.793
80	.869	.861	.892	.853	.894	.836	.817
90	.883	.876	.903	.869	.905	.853	.836
100	.894	.888	.913	.881	.915	.867	.852

Tables of the Constant K (iii)

n	411	321	222	3111	2211	21111	111111
10	.155	.108	.094	.100	.087	.080	.076
11	.192	.140	.123	.130	.114	.106	.099
12	.228	.171	.152	.160	.142	.133	.125
13	.261	.201	.180	.189	.170	.160	.150
14	.292	.230	.208	.217	.197	.186	.176
15	.322	.257	.235	.244	.223	.212	.201
16	.349	.284	.261	.270	.248	.236	.225
17	.375	.309	.285	.295	.272	.260	.248
18	.398	.332	.308	.318	.295	.283	.271
19	.421	.354	.330	.340	.317	.304	.292
20	.442	.375	.351	.361	.338	.325	.313
22	.479	.414	.390	.400	.376	.363	.351
24	.512	.448	.424	.434	.411	.398	.385
26	.542	.479	.456	.465	.442	.430	.417
28	.568	.507	.484	.493	.471	.458	.446
30	.591	.532	.510	.519	.497	.484	.472
35	.640	.585	.564	.573	.552	.540	.528
40	.679	.628	.608	.616	.597	.585	.574
45	.710	.663	.644	.652	.633	.623	.612
50	.736	.692	.674	.681	.664	.654	.644
55	.758	.716	.700	.706	.690	.681	.671
60	.776	.737	.722	.728	.712	.704	.695
65	.792	.755	.740	.746	.732	.723	.715
70	.805	.771	.757	.762	.749	.741	.733
80	.828	.797	.784	.789	.777	.770	.762
90	.846	.818	.806	.811	.800	.793	.786
100	.860	.835	.824	.828	.818	.812	.806

THE MEAN SQUARE SUCCESSIVE DIFFERENCE

By J. von Neumann,[1] R. H. Kent, H. R. Bellinson and B. I. Hart
Aberdeen Proving Ground

1. Introduction. In making measurements, every precaution is generally taken to hold the conditions of the experiment constant, in order that the population, whose parameters are to be estimated from the observations, shall remain fixed throughout the experiment. One wishes each observation to come from the same population, or what is the same thing if normality is assumed, from populations having the same means and standard deviations.

There are cases, however, where the standard deviation may be held constant, but the mean varies from one observation to the next. If no correction is made for such variation of the mean, and the standard deviation is computed from the data in the conventional way, then the estimated standard deviation will tend to be larger than the true population value. When the variation in the mean is gradual, so that a trend (which need not be linear) is shifting the mean of the population, a rather simple method of minimizing the effect of the trend on dispersion is to estimate standard deviation from differences. It is for this purpose that the mean square successive difference

$$\delta^2 = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{n - 1} \tag{1}$$

is suggested. The subscript i in this expression refers to the temporal order of the observation x_i.

In using δ^2 for estimating standard deviation, the distribution of δ^2 in random samples is of interest, since questions of bias, efficiency, and confidence interval require consideration. δ^2 may be used, in addition, to determine whether a trend actually exists; in this case one must know whether δ^2 differs significantly from

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}, \tag{2}$$

which measures variance independently of the order of the observations, and consequently includes the effect of the trend.

---

[1] Institute for Advanced Study, Princeton, N. J. Also member of Scientific Advisory Committee of the Ballistic Research Laboratory, Aberdeen Proving Ground.

The distribution of δ^2 is considered in this paper; it is hoped that others will shortly publish methods of estimating the probability that $\delta^2 \leqq ks^2$ as a function of k and the sample size n.

2. History. A somewhat similar procedure is suggested by "Student" [1] and E. S. Pearson [2] who consider the situation in which a shift may occur in the mean of the population, but where pairs of observations may be made with no shift in mean between them; standard deviation may be estimated from the differences between these pairs. The method can be generalized, and

$$s' = \sqrt{\frac{\sum_{i=1}^{n/2} (x_{2i} - x_{2i-1})^2}{n}}$$

is an estimate of the standard deviation. n must, of course, be an even integer. This estimate has the advantage that its properties are fully known: s' is distributed as the standard deviation with $f = n/2$ degrees of freedom. It will be noted that this estimate does not involve the successive differences, but only the alternate ones. Although there are $n - 1$ available successive differences, this estimate uses only the $n/2$ independent differences. The mean square successive difference is based on all $n - 1$ successive differences, and should therefore provide a more efficient estimate of σ than does s'.

There is, of course, nothing new in the concept of estimating the standard deviation from differences. Even as far back as 1870, an interest in the method appears to have existed. Jordan [3] devised methods based on sums of powers of the differences. Helmert [4] gave more careful consideration to the case of the first power, i.e. the sum of the absolute differences. In both these cases, however, all the $n(n - 1)/2$ differences that can be established from a sample of n observations were included in the estimate, so that the estimate was of no value in reducing the effect of a trend. Helmert realized this, for he pointed out that the estimate obtained from the sum of squares of the differences is exactly that obtained by the more conventional procedure of squaring deviations from the mean.

The usefulness of the differences between successive observations only appears to have been realized first by ballisticians, who faced the problem of minimizing effects due to wind variation, heat and wear in measuring the dispersion of the distance traveled by shell. Vallier [5] appears to have been the first to estimate dispersion from successive differences. Cranz and Becker [6] commended the mean successive difference

$$E_d = \frac{\sum_{i=1}^{n-1} | x_{i+1} - x_i |}{n - 1}.$$

To establish the precision of E_d in estimating σ, Cranz and Becker quoted Helmert's paper, and so erred in saying that their method was superior to that

of the mean deviation. Helmert's procedure, based on $n(n-1)/2$ differences, is indeed more precise (for $n > 10$) than the mean deviation

$$M.D. = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n},$$

but the mean successive difference is based on but $n - 1$ differences, and so is not as precise.

Bennett [7] appears to have suggested the use of successive differences independently of the European ballisticians. In recent years, the method of estimation by the mean square successive difference δ^2 was put into practice in the Ballistic Research Laboratory at the Aberdeen Proving Ground, U. S. Army, by L. S. Dederick

3. Bias and efficiency. The moments of δ^2 in samples drawn from a normal population are derived in Section 6 of this paper. The moments are used at this point to establish the estimate of variance, and the efficiency of this estimate.

The mean value of δ^2 in samples taken at random from a normal population is

$$E(\delta^2) = 2\sigma^2. \tag{3}$$

δ^2 consequently offers an unbiased estimate of variance, and this estimate is

$$\frac{\delta^2}{2} = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{2(n-1)}. \tag{4}$$

The second moment, i.e., the variance, of δ^2 in samples of size n:

$$\sigma_{\delta^2}^2 = \frac{4(3n-4)}{(n-1)^2}\sigma^4. \tag{5}$$

As the sample size is increased, the distribution of δ^2 appears to approach the normal. It is therefore appropriate to consider the efficiency as defined by Fisher [8]. Accordingly, the efficiency of δ^2 is

$$\left[\frac{\sigma_{s^2}}{E(s^2)} \Big/ \frac{\sigma_{\delta^2}}{E(\delta^2)}\right]^2.$$

Since

$$\sigma_{s^2}^2 = \frac{2(n-1)}{n^2}\sigma^4,$$

and

$$E(s^2) = \frac{n-1}{n}\sigma^2,$$

the efficiency of δ^2 in estimating the standard deviation is

$$\frac{2(n-1)}{3n-4} = \frac{2}{3}\left[1 + \frac{1}{3n-4}\right]. \tag{6}$$

The efficiency is unity for $n = 2$, since in this case the two statistics have the same distribution. It therefore appears that the efficiency decreases as the sample size increases, but approaches 2/3 as a limiting value for n very large.

4. Summary of procedure. Having a statistic which estimates a parameter of a population, it is desirable to know the distribution of that statistic as computed from samples taken at random from that population. At present, the distribution of δ^2 in samples of n has not been obtained. The difficulty is in the fact that the successive differences are not independent. The first difference, $d_1 = x_2 - x_1$, and the second difference, $d_2 = x_3 - x_2$, are related in that they both involve x_2. Similar correlation exists between every successive pair of differences between successive observations.

For $n = 2$, and samples taken from a normal population, the distribution of δ^2 is known. Since

$$\delta^2 = (x_2 - x_1)^2 = 2\sum_{i=1}^{2}(x_i - \bar{x})^2 = 4s^2,$$

the distribution of δ^2 is similar to that of s^2 for this sample size.

For $n = 3$, the distribution of δ^2 has been derived analytically. The derivation is indicated in Section 5 of this paper. For $n > 3$, only the moments of the distribution have thus far been obtained. A Pearson type distribution has been fitted to the first three moments to obtain an approximate representation of the true distribution.

5. Distribution of δ^2. In the case of a sample of n taken from a normal population, the probability that the first observation lies between x_1 and $x_1 + dx_1$, while the second lies between x_2 and $x_2 + dx_2$, etc., is

$$\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^n e^{-(x_1^2+x_2^2+\cdots+x_n^2)/2\sigma^2}\, dx_1 dx_2 \cdots dx_n. \tag{7}$$

If $y_i = x_{i+1} - x_i$, this expression becomes

$$\left[\frac{1}{\sigma\sqrt{2\pi}}\right]^n e^{-Q(x_1,y_1,y_2,\cdots,y_{n-1})/2\sigma^2}\, dx_1\, dy_1\, dy_2 \cdots dy_{n-1}, \tag{8}$$

where Q is a quadratic form in x_1 and the y's. Since

$$\delta^2 = \frac{\sum_{i=1}^{n-1} y_i^2}{n-1},$$

the probability that δ^2 shall be less than some value δ_0^2 is

$$(9)\quad P(\delta^2 < \delta_0^2) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^n \underset{\sum_{i=1}^{n-1} y_i^2 < (n-1)\delta_0^2}{\iint \cdots \int} \int_{-\infty}^{+\infty} e^{-Q(x_1, y_1, \cdots, y_{n-1})/2\sigma^2}\, dx_1\, dy_1 \cdots dy_{n-1}.$$

After the integration with respect to x_1 is carried out, the quadratic form in the exponent may be normalized by a transformation to new coordinates z_i linearly related to the y's. The z's may be so chosen that all the terms z_i^2 in the exponent have the same coefficient, in which case

$$(10)\quad P(\delta^2 < \delta_0^2) = c_1 \iint \cdots \int e^{-k\sum_{i=1}^{n-1} z_i^2} \frac{\partial(y_1, y_2, \cdots, y_{n-1})}{\partial(z_1, z_2, \cdots, z_{n-1})}\, dz_1\, dz_2 \cdots dz_{n-1}.$$

As a result of such a transformation, the sphere of integration in (9) becomes an ellipsoid in (10). By changing to polar coordinates, with

$$r^2 = \sum_{i=1}^{n-1} z_i^2,$$

$$(11)\quad P(\delta^2 < \delta_0^2) = c_1 \iint e^{-kr^2} r^{n-2}\, d\Omega\, dr,$$

in which Ω is the solid angle in the space of $n - 1$ dimensions. The limits of integration with respect to Ω as a function of r must be found; this involves the evaluation of the solid angle subtended by the surface bounded by the intersection of the $(n - 1)$-dimensional sphere and the $(n - 1)$-dimensional ellipsoid. If $\Omega = \phi(r)$,

$$(12)\quad P(\delta^2 < \delta_0^2) = c_2 \int_0^a e^{-kr^2} \phi(r) r^{n-2}\, dr,$$

in which a is the longest semi-axis of the $(n - 1)$-dimensional ellipsoid corresponding to the given value of δ^2.

For $n = 3$, (9) becomes

$$\begin{aligned}(13)\quad P(\delta^2 < \delta_0^2) &= \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^3 \underset{y_1^2 + y_2^2 < 2\delta_0^2}{\iint} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{3\sigma^2}(y_1^2 + y_2^2 + y_1 y_2)\right. \\ &\qquad \left. - \frac{3}{2\sigma^2}\left(x_1 + \frac{2y_1 + y_2}{3}\right)^2\right] dx_1\, dy_1\, dy_2 \\ &= \frac{1}{2\sqrt{3}\,\pi\sigma^2} \underset{y_1^2 + y_2^2 < 2\delta_0^2}{\iint} e^{-(y_1^2 + y_1 y_2 + y_2^2)/3\sigma^2}\, dy_1\, dy_2.\end{aligned}$$

Normalizing the quadratic form in the exponent,

$$(14)\quad P(\delta^2 < \delta_0^2) = \frac{1}{2\sqrt{3}\,\pi\sigma^2} \underset{z_1^2 + z_2^2 < 2\delta_0^2}{\iint} e^{-(z_1^2 + \frac{1}{3}z_2^2)/2\sigma^2}\, dz_1\, dz_2,$$

and in polar coordinates

$$
(15)\quad \begin{aligned}
P(\delta^2 < \delta_0^2) &= \frac{1}{2\sqrt{3}\,\pi\sigma^2}\int_0^{\delta_0\sqrt{2}}\int_0^{2\pi} re^{-r^2[\cos^2\theta+\frac{1}{3}\sin^2\theta]/2\sigma^2}\,d\theta\,dr \\
&= \frac{1}{2\sqrt{3}\,\pi\sigma^2}\int_0^{\delta_0\sqrt{2}} re^{-r^2/2\sigma^2}\left[\int_0^{2\pi} e^{r^2\sin^2\theta/3\sigma^2}\,d\theta\right]dr.
\end{aligned}
$$

The integral in brackets can be shown to be a Bessel function of zero order; for let

$$r^2/3\sigma^2 = -2iu,$$

$$\phi = \frac{\pi}{2} - 2\theta,$$

then

$$(16)\quad \int_0^{2\pi} e^{r^2\sin^2\theta/3\sigma^2}\,d\theta = e^{-iu}\int_{-\pi}^{\pi} e^{iu\sin\phi}\,d\phi = 2\pi e^{-iu}J_0(u).$$

Consequently, (15) takes the form

$$(17)\quad P(\delta^2 < \delta_0^2) = \frac{1}{\sigma^2\sqrt{3}}\int_0^{\delta_0\sqrt{2}} re^{-r^2/3\sigma^2}J_0\left(\frac{ir^2}{6\sigma^2}\right)dr = F(\delta_0^2).$$

The probability density function

$$
(18)\quad \begin{aligned}
p(\delta^2) &= \frac{dF(\delta^2)}{d\delta^2} \\
&= \frac{1}{\sigma^2\sqrt{3}}e^{-2\delta^2/3\sigma^2}J_0\left(\frac{i\delta^2}{3\sigma^2}\right) \\
&= \frac{1}{\sigma^2\sqrt{3}}e^{-2\delta^2/3\sigma^2}\left[1 + \frac{1}{2^2}\frac{\delta^4}{3^2\sigma^4} + \frac{1}{2^24^2}\frac{\delta^8}{3^4\sigma^8} + \frac{1}{2^24^26^2}\frac{\delta^{12}}{3^6\sigma^{12}} + \cdots\right].
\end{aligned}
$$

6. Moments. The t-th moment of δ^2 about the origin is defined by

$$(19)\quad \mu_t' = E[(\delta^2)^t],$$

or

$$
(20)\quad \begin{aligned}
(n-1)^t\mu_t' &= E\left(\left[\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2\right]^t\right) \\
&= E\left(\left[2\sum_{i=1}^{n} x_i^2 - (x_1^2 + x_n^2) - 2\sum_{i=1}^{n-1} x_{i+1}x_i\right]^t\right).
\end{aligned}
$$

For any value of t, the expansion can be performed, and similar terms collected and enumerated. The values of x can be considered as true errors, i.e. as deviations from the true mean, without affecting the conclusions. If the

original population from which the samples have been drawn is normal, with standard deviation σ, then:

$$E(x^{2k-1}) = 0$$

(21)

$$E(x^{2k}) = \frac{(2k)!}{2^k k!}\sigma^{2k},$$

and since, in the null case where the mean of the population remains constant, successive observations are independent, then

(22)

$$E(x_i^r x_j^s) = E(x^{r+s}), \qquad i = j$$

$$E(x_i^r x_j^s) = E(x^r)E(x^s), \qquad i \neq j.$$

These relations are sufficient for the evaluation of μ_t'. For example, in the case of the second moment, $t = 2$:

(23) $$(n-1)^2\mu_2' = E\left(\left[2\sum_{i=1}^{n} x_i^2 - (x_1^2 + x_n^2) - 2\sum_{i=1}^{n-1} x_{i+1}x_i\right]^2\right).$$

Now:

$$\left[2\sum_{i=1}^{n} x_i^2 - (x_1^2 + x_n^2) - 2\sum_{i=1}^{n-1} x_{i+1}x_i\right]^2$$

$$= 4\left(\sum_{i=1}^{n} x_i^2\right)^2 + (x_1^2 + x_n^2)^2 + 4\left(\sum_{i=1}^{n-1} x_{i+1}x_i\right)^2$$

$$- 4(x_1^2 + x_n^2)\sum_{i=1}^{n} x_i^2 - 8\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n-1} x_{i+1}x_i + 4(x_1^2 + x_n^2)\sum_{i=1}^{n-1} x_{i+1}x_i$$

$$= 4\left[\sum_{i=1}^{n} x_i^4 + \sum_{i,j=1,i\neq j}^{n} x_i^2x_j^2\right] + [x_1^4 + 2x_1^2x_n^2 + x_n^4]$$

$$+ 4\left[\sum_{i=1}^{n-1} x_{i+1}^2x_i^2\right] - 4\left[x_1^4 + x_1^2\sum_{i=2}^{n} x_i^2 + x_n^2\sum_{i=1}^{n-1} x_i^2 + x_n^4\right]$$

$$+ \text{[terms containing odd powers of } x_i\text{]}.$$

The mean of these terms is found by using (21) and (22), and the number of each type of term present is enumerated:

$$4[n(3\sigma^4) + n(n-1)\sigma^2\sigma^2] + [3\sigma^4 + 2\sigma^2\sigma^2 + 3\sigma^4] + 4[(n-1)\sigma^2\sigma^2]$$

$$- 4[3\sigma^4 + \sigma^2(n-1)\sigma^2 + \sigma^2(n-1)\sigma^2 + 3\sigma^4] = (4n^2 + 4n - 12)\sigma^4.$$

Consequently

(24) $$\mu_2' = \frac{4(n^2 + n - 3)}{(n-1)^2}\sigma^4.$$

The first four moments about the origin were evaluated by this procedure,

and from these, the moments about the mean are readily determined. The results are:

$$\mu_1' = 2\sigma^2$$

$$\mu_2' = \frac{4(n^2 + n - 3)}{(n-1)^2}\sigma^4$$

$$\mu_3' = \frac{8(n^3 + 6n^2 + 2n - 21)}{(n-1)^3}\sigma^6$$

$$\mu_4' = \frac{16(n^4 + 14n^3 + 53n^2 - 8n - 231)}{(n-1)^4}\sigma^8 \tag{25}$$

$$\mu_1 = 0$$

$$\mu_2 = \frac{4(3n - 4)}{(n-1)^2}\sigma^4$$

$$\mu_3 = \frac{32(5n - 8)}{(n-1)^3}\sigma^6$$

$$\mu_4 = \frac{48(9n^2 + 46n - 112)}{(n-1)^4}\sigma^8.$$

It should be noted at this point that the above fourth moment is incorrect for $n = 2$. One of the terms in the expansion of the right side of (20), for $t = 4$, is

$$x_1^2 x_n^2 \sum_{i=1}^{n-1} x_{i+1}^2 x_i^2.$$

For $n = 2$, the mean value of this term is

$$E(x_1^2 x_2^2 x_2^2 x_1^2) = E(x_1^4)E(x_2^4) = 9\sigma^8,$$

whereas for $n > 2$, the mean value is

$$E(x_1^4 x_2^2 x_n^2) + E\left(x_1^2 x_n^2 \sum_{i=2}^{n-2} x_{i+1}^2 x_i^2\right) + E(x_1^2 x_{n-1}^2 x_n^4) = (n+3)\sigma^8.$$

7. Pearson type fit to distribution of δ^2. From the moments it is found that

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{16(5n-8)^2}{(3n-4)^3},$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3(9n^2 + 46n - 112)}{(3n-4)^2}. \tag{26}$$

As n becomes large, β_1 and β_2 approach 0 and 3 respectively; the distribution therefore appears to approach the normal for large samples. For finite sample sizes, the values of β_1 and β_2 correspond to those of the Pearson Type VI

distribution,

$$p\left(\frac{\delta^2}{\sigma^2}\right) = c\left(\frac{\delta^2}{\sigma^2} + a_1\right)^{q_2}\left(\frac{\delta^2}{\sigma^2} + a_2\right)^{-q_1}.$$

The origin of this distribution is at $\delta^2 = -a_1\sigma^2$, but the origin of the true distribution must be at $\delta^2 = 0$. By taking $a_1 = 0$ so that the origin is at $\delta^2 = 0$, we obtain what appears to be a suitable approximation

$$p\left(\frac{\delta^2}{\sigma^2}\right) = c\left(\frac{\delta^2}{\sigma^2}\right)^{q_2}\left(\frac{\delta^2}{\sigma^2} + a_2\right)^{-q_1}. \tag{27}$$

The parameters are determined by equating the 1st, 2nd and 3rd moments of (27) to the corresponding moments of the true distribution, with the result that

$$\begin{aligned} q_2 &= \frac{3n^4 - 10n^3 - 18n^2 + 79n - 60}{8n^3 - 50n + 48}, \\ q_1 &= \frac{4 - \mu_2(q_2 + 1)(q_2 + 3)}{4 - \mu_2(q_2 + 1)}, \\ a_2 &= \frac{2(q_1 - q_2 - 2)}{q_2 + 1}, \\ c &= \frac{a_2^{q_1 - q_2 - 1}}{\mathrm{B}(q_2 + 1, q_1 - q_2 - 1)}. \end{aligned} \tag{28}$$

Values of these parameters for selected values of n are given in Table I. The sixth and seventh columns of this table give the values of β_2 for the distribution (27) and for the true distribution, respectively.

TABLE I

(1) n	(2) q_1	(3) q_2	(4) a_2	(5) c	(6) β_2 (27)	(7) β_2 True	(8) Ratio (6)/(7)
5	24.4391	0.6391	26.6000	5.8800×10^{34}	8.807	8.504	1.036
7	31.1286	1.3857	23.2571	4.9285×10^{42}	6.948	6.758	1.028
10	41.2830	2.5079	20.9667	9.4934×10^{54}	5.658	5.538	1.022
15	58.2113	4.3806	19.2659	4.0240×10^{75}	4.718	4.645	1.016
20	75.1210	6.2543	18.4351	1.8063×10^{96}	4.269	4.217	1.012
25	92.0189	8.1285	17.9417	8.1097×10^{116}	4.006	3.965	1.010
50	176.4443	17.5018	16.9651	1.3386×10^{220}	3.494	3.475	1.005

The *Tables of the Incomplete Beta-Function* [9] can be used to evaluate the probability integral of the distribution (27),

$$\begin{aligned} P\left(\frac{\delta^2}{\sigma^2} < \frac{\delta_0^2}{\sigma^2}\right) &= c\int_0^{\delta_0^2/\sigma^2}\left(\frac{\delta^2}{\sigma^2}\right)^{q_2}\left(\frac{\delta^2}{\sigma^2} + a_2\right)^{-q_1} d\left(\frac{\delta^2}{\sigma^2}\right) \\ &= 1 - I_x(q_1 - q_2 - 1, q_2 + 1) \\ x &= \frac{a_2}{a_2 + \delta_0^2/\sigma^2}, \end{aligned} \tag{29}$$

for $n \leqq 14$. For $n > 14$, the probability integral may be determined by quadrature. Some values of the probability integral for $n = 50$ are given in Table II. A comparison with the integral of the normal curve having the same first two moments indicates that a sample of somewhat more than 50 is required before the normal curve becomes a satisfactory approximation to the distribution (27).

TABLE II

$$P\left(\frac{\delta^2}{\sigma^2} < \frac{\delta_0^2}{\sigma^2}\right) \quad \textit{for } n = 50$$

δ_0^2/σ^2	(29)	Normal
.50	.00000	.00118
.75	.00031	.00563
1.00	.00647	.02129
1.25	.04393	.06418

REFERENCES

[1] "Student," *Biometrika*, Vol. 19(1927), p. 158.

[2] E. S. Pearson, *Application of Statistical Methods to Industrial Standardisation and Quality Control*, London, 1935, p. 62.

[3] W. Jordan, *Astronomische Nachrichten*, Vol. 74(1869), pp. 209–226.

[4] F. R. Helmert, *Astronomische Nachrichten*, Vol. 88(1876), pp. 127–132.

[5] E. Vallier, *Balistique Experimentale*, Paris, 1894, p. 166.

[6] C. Cranz and K. Becker, *Exterior Ballistics*, (trans. from 2nd German edition) London, 1921, p. 383.

[7] A. A. Bennett, unpublished report to the Chief of Ordnance, U. S. Army, circa 1918.

[8] R. A. Fisher, *Phil. Trans.* A, Vol. 222(1922), p. 316.

[9] Karl Pearson (Editor), *Tables of the Incomplete Beta-Function*, London: Biometrika Office, 1934.

THE RETURN PERIOD OF FLOOD FLOWS

By E. J. Gumbel

New School for Social Research

Introduction. Engineers have used various interpolation formulas to represent the observed distribution of flood discharges. These formulas are sometimes constructed *ad hoc* for a given stream, and have no general meaning. Most of them are rather complicated.[1] Some authors have tried to introduce upper and lower limits to the discharges, even though it is doubtful that such limits exist. Others have introduced the third and fourth moments of the distribution, in spite of the fact that these numerical values are subject to large errors. For some formulas it is impossible to give a meaning to the constants; different formulas applied to the same stream give rather contradictory results; and consequently there is considerable confusion. For example, Slade [20] has stated that "the statistical method in whatever form employed is an entirely inadequate tool in the determination of flood frequencies." According to Saville [19] "the engineer should satisfy himself that he has used an adequate number of methods, whether mathematical, graphic or otherwise, which have real support from either theory or experience, and then form his own judgement."

The main reason for this situation is that these studies have little or no theoretical basis. The author believes it possible to give exact solutions, exactitude being interpreted from the standpoint of the calculus of probabilities [10]. Our solutions are simply the consequences of a truism: "The flood discharges are the largest values of the discharges." The present study is but an explanation of this statement.

Many American authors start with a statistical function, which we call the return period of floods. Therefore we shall first analyse the notion of return period and show how it can be derived as a consequence of the concept of distribution. We then give a short résumé of the theory of largest values. The discharge, and in consequence the flood discharge, is considered as an unlimited statistical variable; it is not necessary to determine its distribution. We are justified in representing the observed distribution of flows by one of the theoretical distributions of largest values. The distribution we choose contains only two constants, and both have a clear hydrological meaning. The numerical values are calculated by the method of moments.

---

[1] In recent years many articles discussing this topic have been published by the American Society of Civil Engineers and the American Geophysical Union [8]. A review of some of the proposed formulas is given in the Water Supply Paper 771 [17].

The application of the notion of return period to the largest values leads to a simple formula for the return period of the floods. In the last part of this paper we represent the flood flows of the Rhône and Mississippi Rivers by our formula.

1. The return period. Let us consider a continuous statistical variable x, having a theoretical distribution $w(x)$. The probability $W(x)$ of a value less than or equal to x, and the probability $P(x)$ of a value greater than or equal to x, are

$$W(x) = \int_{-\infty}^{x} w(z)\,dz, \qquad P(x) = \int_{x}^{\infty} w(z)\,dz, \tag{1}$$

where z denotes the variable of integration. Clearly

$$W(x) + P(x) = 1. \tag{1'}$$

Let n be the number of observations. Let x_m $(m = 1, 2, \cdots, n)$ be the observed values arranged in increasing magnitude, where m is the serial number beginning with the lowest ("from below"). The lowest observation has the serial number $m = 1$, the highest has the serial number $m = n$. These observed values will be written x_1, and x_n respectively. The number of observations below or equal to x_m is $m = n'W\,(x_m)$ where $'W(x_m)$ is the observed relative number corresponding to the probability $W(x)$. The graphic representation of this series is called a cumulative histogram.

In hydraulics many authors arrange the observations in decreasing magnitude. Let ${}_mx$ $(m = 1, 2, \cdots, n)$ be these observed values. The serial number m is counted in a descending scale ("from above"). For the largest value $m = 1$, for the lowest value $m = n$. The number of observations above or equal to ${}_mx$ is $m = n'P({}_mx)$ where $'P({}_mx)$ corresponds to $P(x)$. The numbers $'W(x_m)$ will never decrease; the number $'P({}_mx)$ will never increase. The mth value on a descending scale is the $n - m + 1$th value on an ascending scale. Therefore

$$n'P({}_mx) = n - n'W(x_m) + 1, \tag{2}$$

and

$$nP(x) = n - nW(x). \tag{2'}$$

The difference between formulas (2) and (2') will play a certain rôle later.

Different methods are used in statistics in comparing the theoretical values $W(x)$ or $P(x)$ and $w(x)$ with the corresponding observations $'W(x_m)$, or $'P({}_mx)$ (cumulative frequencies) and $\Delta'W(x_m)$ (frequency distribution). They all have in common an arrangement of observed values according to magnitude.

For the purpose of considering the observations in chronological order, we introduce a statistical criterion which at first glance may appear to have a new logical structure. It is assumed here that the observations are made at constant time intervals, and this interval is considered the unit of time. We suppose that the observations are homogeneous, i.e., subject to a common set of forces.

Furthermore, we suppose that the events are independent of one another: the occurrence of a high or low value for x has no influence on the value of any succeeding observation. Let us choose a low value x, and ask the following: After what number of observations does this or a greater value return? We calculate the mean of these chronological intervals between every two consecutive values, equal to or greater than x. We repeat these operations for a second, third, . . . till the penultimate value of x.

These means are called the *observed return periods*. The criterion consists of the comparison of the observed, and the theoretical return period for increasing values of x. For a discontinuous variable we could obtain the return period for a value equal to x, (not equal to or greater than x). This average time, which is sometimes used in physics, does not interest us, as our variable, the discharge, is continuous. We limit our consideration to the return period of a value equal to or greater than x, called: value greater than x.

The determination of the theoretical return period is a classical problem: How many trials must, on the average, be made, in order that an event of a given probability should happen? Our event, the realization of a value, equal to or greater than x, has the probability $P(x) = 1 - W(x)$.

The mean number of trials $T(x)$ which are necessary to obtain our event once, is evidently

$$T(x) = \frac{1}{1 - W(x)}, \tag{3}$$

or

$$T(x) = \frac{1}{P(x)}. \tag{3'}$$

This value $T(x)$ is the mean chronological interval between two values, equal to or greater than x. If we start at the time when such a value has been observed for the first time, we can interpret $T(x)$ as the theoretical return period of a value equal to or greater than x. We designate it as the *theoretical return period*. This concept has not been used in statistics. It is a well-known concept in hydraulics which was introduced by Fuller [6]. To every theoretical distribution $w(x)$ there is a corresponding return period $T(x)$ and conversely, to every theoretical return period $T(x)$ there is a corresponding distribution

$$w(x) = \frac{T'(x)}{T^2(x)}, \tag{4}$$

obtained by differentiating (3).

If the variable is without limit to the left, the return period will start with $T = 1$. If the variable is limited to the left by $x \geqq \epsilon$ the corresponding return period will be

$$T(\epsilon) \geqq 1 \qquad \text{if } W(\epsilon) \geqq 0 \tag{5}$$

In the graphic representation, the return period $T(x)$ which has a time dimension, will be the abcissa and x the ordinate. Therefore we consider x as a function of $T(x)$; from (4) we obtain

$$\frac{dx}{d \ln T} = \frac{1}{w(x)T(x)} \tag{6}$$

where ln signifies the natural logarithm. The increase of x as a function of $\ln T(x)$ will be very rapid for small values of T. For a limited distribution the same result is obtained, provided the probability $W(\epsilon)$ and the density of probability $w(\epsilon)$ are sufficiently small. Clearly, the return periods of the three quartiles are respectively $1\frac{1}{3}$, 2, 4. The return period will always increase with x. It will tend towards infinity even if the variable is limited to the right.

Let us now consider the calculus of the observed return periods. Instead of values equal to or greater than x_m we will only speak of values greater than x_m. The observed return period is the interval between the first and the last observation greater than x_m, divided by the number of intervals between all observations greater than x_m. The number of observations greater than x_m is $n - n'W(x_m)$. Between these observations there are $n - n'W(x_m) - 1$ intervals. This denominator is independent of the chronological order of the observed values. We can calculate the mean of the observed intervals up to a value x_m so that $n - n'W(x_m) = 2$. For this value of x_m there are only two observations, i.e., only one interval. In that case no mean can be calculated.

The numerator, the interval between the first and the last observation greater than x_m will be $n - 1$, provided that the first and the last value in chronological order are greater than x_m. But in general the first value greater than x_m will be the $('k + 1)$th in chronological order. The first value greater than x_m found in the reverse chronological order, will be the $(k' + 1)$th. Let $'k + k' = l$, then the interval between the last and the first value greater than x_m is $n - 1 - l$. The mean observed interval is thus

$${}_1T(x_m) = (n - 1 - l)/(n - 1 - n'W(x_m)),$$

or

$${}_1T(x_m) = \left(1 - \frac{l}{n-1}\right) \Big/ \left(1 - \frac{m}{n-1}\right). \tag{7}$$

This magnitude depends only on the chronological order of the first and the last value greater than x_m. It is independent of the chronological order of all other observations. Even in the case $l = 0$ this value differs from the theoretical value (3). The observed value surpasses the theoretical value, even if the frequency $'W(x_m)$ is identical with the probability $W(x)$.

In the general case, $l > 0$, this difference is a function of l. The number l depends upon the times at which the observations begin and cease; but it is not a characteristic of the chronological order. As a result of these disadvantages of formula (7) we prefer to introduce other definitions, in which the

chronological order does not enter. These definitions have an added advantage in that they are constructed in a manner analogous to the theoretical formula.

The observed value which corresponds to (3) is

$$'T(x_m) = \frac{n}{n - n'W(x_m)}, \tag{8}$$

or

$$'T(x_m) = n/(n - m). \tag{9}$$

But this definition of the observed return period is not the only one which corresponds to (3). Starting with the serial number m, in a descending scale, Fuller [6] puts

$$''T(x_m) = \frac{n}{m}. \tag{8'}$$

According to this definition, the return period of the mth value from below is

$$''T(x_m) = n/(n - m + 1). \tag{9'}$$

TABLE I

Two definitions of the observed return periods

observed variable	serial number from below	serial number from above	exceedance interval formula (9)	recurrence interval formula (9′)
x_1	1	n	$n/(n - 1)$	1
x_2	2	$n - 1$	$n/(n - 2)$	$n/(n - 1)$
x_m	m	$n - m + 1$	$n/(n - m)$	$n/(n - m + 1)$
x_{n-1}	$n - 1$	2	$n/1$	$n/2$
x_n	n	1	—	$n/1$

This observed return period corresponds to the theoretical return period (3′). The difference between (9) and (9′) results from the fact that the relation (2) between the observed cumulative frequencies $'W(x_m)$ and $'P(_mx)$ differs from the relation (2′) between the probabilities $W(x)$ and $P(x)$. The two definitions of the observed return periods are related by

$$''T(x_{m+1}) = 'T(x_m) < 'T(x_{m+1}). \tag{10}$$

From a purely logical standpoint the first definition is as justifiable as the second one.. Both are used in hydraulics. In order to avoid confusion between formulas (9) and (9′) Horton [16] calls $'T(x_m)$ the *exceedance* interval, i.e., "the average interval at which an event of given magnitude is exceeded," whereas he defines $''T(x_m)$, the recurrence interval as "the average interval of occurrence of values equalling or exceeding a given magnitude." Of course, the exceedance interval surpasses the recurrence interval. Since both observed intervals correspond to a common theoretical return period we designate both of them as observed return periods.

The difference between formulas (9) and (9′) is made clear in Table I.

Each of the definitions (9) and (9′) and the theoretical expression $T(x)$ has different properties. For the lowest observation

$$n'W(x_1) = 1; \qquad n'P({}_nx) = n.$$

Therefore

$$'T(x_1) = 1 + \frac{1}{n-1}; \qquad ''T(x_1) = 1,$$

whereas for an unlimited distribution $\lim_{x \to -\infty} T(x) = 1$.

If the number of observations is sufficiently large the numerical differences between the two observed periods are rather small, except for very large values of the variable. For the last observation

$$n'W(x_n) = n; \qquad n'P({}_1x) = 1.$$

Therefore the return period $'T(x_n)$ for the last observation does not exist. According to the second definition the return period for the last value is equal to the total number of observations. But in general there is only one observation of the last value.

The preference given formula (9) over (9′) corresponds with the preference given to $W(x)$ over $P(x)$ when comparing the theoretical with the observed values. Therefore it is natural to count m from below. Since both definitions are equally applicable and since they lead to different results for large values of the variable, one should not calculate the return period for a small number of observations.

The observed return periods (9) and (9′) differ from the theoretical return period (3) in the same way that the frequencies $'W(x_m)$ or $'P({}_mx)$ differ from the probabilities $W(x)$ or $P(x)$. The chronological order enters neither into formula (7) nor into (9) or (9′). We need not take it into consideration, since the theoretical return period is obtained from the probability and the observed return period from the cumulative histogram. Therefore the usual statistical methods can be used for making the comparison between observed and theoretical return periods.

The return period is a statistical function like the distribution $w(x)$ or the probability $W(x)$. No formula for $T(x)$ that contradicts the properties of $w(x)$ can be accepted. The return period $T(x)$ will contain the same number of independent constants as the distribution $w(x)$. Consequently the fit of the theoretical curve $T(x)$ to the observations $'T(x_m)$ or $''T(x_m)$ cannot be improved by introducing a new constant without also changing the distribution $w(x)$. The theoretical curve $x = f(T)$ will fit the observed curves $(x_m, 'T(x_m))$ and $(x_m, ''T(x_m))$ in a way that depends upon the fit of $W(x)$ and $P(x)$ to $'W(x_m)$ and $'P({}_mx)$.

Let us suppose that $w(x)$ contains k constants; that they are determined by the method of moments which conserves the arithmetic mean $\bar{x}$, the mean of the squares $\overline{x^2}$ etc. of the observed distribution. For the return period these mo-

ments have a meaning. Let us consider for the sake of simplicity a positive variable. The kth moment M_k

$$M_k = \int_0^\infty x^k \, dW(x)$$

$$= -\int_0^\infty x^k \, d(1 - W(x))$$

$$= k \int_0^\infty (1 - W(x)) x^{k-1} \, dx$$

is according to (3)

(11) $$M_k = k \int_0^\infty \frac{x^{k-1} \, dx}{T(x)},$$

whence for $k = 1$ and $k = 2$

(11′) $$\bar{x} = \int_0^\infty \frac{dx}{T(x)}; \qquad E(x^2) = 2 \int_0^\infty \frac{x \, dx}{T(x)}.$$

For a given distribution containing two constants, the method of moments conserves the area and the center of gravity of the reciprocal of the return period. Even if the method of methods moments gives the best determination of the constants, for the distribution, it need not give the best determination for the return period. But if the observed return periods were used for the determination of the constants we would get two sets, since there are two observed curves having equal validity, but different values for large x. We will get one and only one set if the constants are calculated from the observed distribution, for here the difference between $'T(x_m)$ and $''T(x_m)$ does not matter. The fact that we do not take the constants from the observed return periods, but from another statistical function, might be a cause for deviations between the observed and the theoretical return periods.

Once the constants have been found, we compare the observed curves $(x_m, 'T(x_m))$ and $(x_m, ''T(x_m))$ with the theoretical curve $x = f(T)$. To avoid discontinuity the observed return period will be established for all values of x_m arranged in increasing order.

If the observed return periods for small values of x are systematically smaller (greater) than the theoretical period, it is reasonable to conclude that there exists an attraction (repulsion) for small values of the variable and a repulsion (attraction) for the large values. But it must be remembered that the observed values have different weights in that the return periods for small values of x are based on many observations. This number diminishes as x increases. The last observed return period is based only on two observations. Therefore the divergence between theory and observation will increase with the variable. With this precaution the criterion of the return period suggests one cause of difference between theory and observation. In order to apply this method to the largest values we must first establish the corresponding distribution.

2. Theory of the largest value. Let x be a statistical variable unlimited to the right having the distribution $w(x)$. Among the N observed values, one will be larger than the others. We wish to determine its theoretical value.

According to the principle of multiplication the probability $\mathfrak{W}_N(x)$ that N values are inferior to x is

$$\mathfrak{W}_N(x) = W^N(x). \tag{12}$$

This is the probability of x being the largest value. The largest value is a new statistical variable which possesses a mode, a mean $\bar{u}$, a standard deviation s and higher moments. To get the mean the distribution $\mathfrak{w}_N(x)$ of the largest value is needed. From (12) by differentiation

$$\mathfrak{w}_N(x) = NW^{N-1}(x)w(x). \tag{13}$$

The mode will be the solution of

$$\frac{N-1}{W(x)}\, w(x) + \frac{w'(x)}{w(x)} = 0. \tag{13'}$$

For a given initial distribution $w(x)$ and for small N we have to solve this equation. But the mean and the moments cannot be obtained in a general way by the use of the exact distribution (13). However we can reach general solutions if N is large, provided we limit ourselves to certain classes of initial distributions. We have studied this problem in previous publications [11–13]. For our present purpose it is sufficient to give the results in a form due to R. von Mises [18].

We define a large value u of the variable x by

$$N(1 - W(u)) = 1. \tag{14}$$

This means that the expected number of observations equal to or greater than u is one. Equation (14) is but another form of definition (3). The mean number of trials is used in (3) whereas the original variable x is used in (14).

The probability $\alpha\, du$ that a value greater than u will be contained between u and $u + du$ is given by

$$\alpha = \frac{w(u)}{1 - W(u)}. \tag{15}$$

Obviously α and u are functions of N and the constants in the initial distribution $w(x)$. There are two limiting forms of the probability (12)

$$\lim_{N\to\infty} W^N(x) = F(x); \qquad \lim_{N\to\infty} W^N(x) = \mathfrak{W}(x).$$

If

$$\lim_{u\to\infty} \alpha u = k > 0, \tag{16}$$

we obtain

$$F(x) = e^{-(u/x)^k}. \tag{17}$$

This probability function was first established by Fréchet [5]. If

$$\lim_{u\to\infty} \frac{d}{du}\left(\frac{1}{\alpha}\right) = 0, \tag{18}$$

we obtain

$$\mathfrak{W}(x) = e^{-e^{-\alpha(x-u)}}. \tag{19}$$

This probability function is due to R. A. Fisher [4]. Let us consider the first limit. The initial distributions which lead to it belong to the *Pareto type.* For this distribution

$$w(x) = \frac{k}{x^{k+1}}; \qquad W(x) = 1 - \frac{1}{x^k}; \qquad x \geqq 1$$

and condition (16) holds; for *any* value of x

$$\frac{xw(x)}{1 - W(x)} = k.$$

The distribution $f(x)$ of the largest value, which corresponds to (17), is

$$f(x) = \frac{k}{u}\left(\frac{u}{x}\right)^{k+1} e^{-(u/x)^k}. \tag{20}$$

The mode $\tilde{x}_N$ of the largest value is the solution of

$$\frac{d}{dx}\left[(k+1)\ln\frac{u}{x} - \left(\frac{u}{x}\right)^k\right] = 0,$$

hence

$$\frac{k+1}{x} = \frac{ku^k}{x^{k+1}},$$

or

$$\tilde{x}_N = u\left(\frac{k}{k+1}\right)^{1/k}. \tag{21}$$

According to the definition (14) the mode of the largest value will increase with N. For a finite number of observations, which is always the case, the mode will be limited. But the moments of order k or higher will not exist. For $k < 1$, no moment will exist. For $k < 2$, only the first moment, the mean, exists, and so on.

Let us consider now the second limit (19). The initial distributions which lead to it belong to the *exponential type*. For this distribution [14]

$$w(x) = e^{-x}; \qquad W(x) = 1 - e^{-x}; \qquad x \geqq 0,$$

and for *any* value of x

$$\frac{d}{dx}\left(\frac{1 - W(x)}{w(x)}\right) = 0,$$

which means that condition (18) is fulfilled. Most of the distributions used in statistics belong to this type. According to (19) the distribution of the largest value is

$$\mathfrak{w}(x) = \alpha e^{-\alpha(x-u)-e^{-\alpha(x-u)}}. \tag{22}$$

If we introduce a reduced variable y without dimension by the linear transformation

$$y = \alpha(x - u), \tag{23}$$

we get the reduced probability $\mathfrak{V}(y)$

$$\begin{aligned} \mathfrak{V}(y) &= \mathfrak{W}(x) \\ &= e^{-e^{-y}}. \end{aligned} \tag{24}$$

The numerical values of this function, calculated by means of Becker's tables [1], are given in Table II, col. 1 and 2. The reduced distribution

$$\mathfrak{v}(y) = e^{-y-e^{-y}}, \tag{25}$$

makes clear the meaning of u: the distribution has one and only one maximum which occurs for the reduced value $y = 0$. Therefore u is the mode of the largest value for a given set of N observations. For an initial distribution $w(x)$ satisfying (18), and for large N, definition (3) of the return period as a function of x becomes identical with relation (14) which involves the number of observations N and the corresponding most probable value u.

We wish to decide which distribution of the largest value is to be used to represent the given observations. This decision depends, according to (16) and (18), on the nature of the initial distribution at the extreme values of the variable. If the law of the observed initial variable is known, a precise answer can be given. But generally speaking, a distribution chosen to represent given observations is nothing but an interpolation formula. Formulas having different analytical properties may all give satisfactory results. One might fulfill condition (16), and another (18). The conditions apply to the differential coefficient, whereas the initial observations are always discontinuous. Therefore they will not enable us to decide which, if any, of the conditions is met. For extreme values of the variable x the observed differences are large and nonuniform, and there is therefore no way to replace the differentiation by a finite difference. Consequently we have to use the observations of the largest values to control the two competing theories and not the conditions. The fact that distribution (20) has higher moments only under certain conditions, is a strong practical argument in favor of distribution (22). Therefore the following development will be based on this distribution.

It can be shown that the mean error θ of distribution (22) is related to the constant α by

$$\theta = 0.98/\alpha. \tag{26}$$

Therefore the constant u is the most probable largest value for N observations and $1/\alpha$ a multiple of the mean error.

TABLE II

Probabilities and return periods of largest values

reduced variable y	probability $\mathfrak{W}(x)$	return period $\log T(x)$	Flood discharges per second	
			in cubic meter x Rhône R.	in 1000 cubic feet x Mississippi R.
−2.00	0.00062	0.000		
−1.75	0.00317	0.001		
−1.50	0.01131	0.005	1355	803
−1.25	0.03049	0.013	1492	869
−1.00	0.06599	0.030	1629	936
−0.75	0.12039	0.056	1766	1002
−0.50	0.19230	0.093	1903	1069
−0.25	0.27693	0.141	2040	1135
0.00	0.36788	0.199	2177	1202
0.25	0.45896	0.267	2314	1268
0.50	0.54524	0.342	2451	1335
0.75	0.62352	0.424	2588	1401
1.00	0.69220	0.512	2725	1468
1.25	0.75088	0.604	2862	1534
1.50	0.80001	0.699	2999	1601
1.75	0.84048	0.797	3136	1667
2.00	0.87342	0.899	3273	1734
2.25	0.89996	1.000	3410	1800
2.50	0.92119	1.103	3547	1867
2.75	0.93807	1.208	3686	1933
3.00	0.95143	1.314	3822	2000
3.25	0.96197	1.420	3959	2066
3.50	0.97025	1.527	4096	2133
3.75	0.97675	1.634	4233	2199
4.00	0.98185	1.741	4370	2266
4.25	0.98584			
4.50	0.98895			
4.75	0.99138			
5.00	0.99329			
5.25	0.99477			
5.50	0.99592			
5.75	0.99682			
6.00	0.99752			

TABLE III

Observed return periods

Rhône, Lyon (France) (1826–1936)

Flood discharge x_m	Serial number m	Return period $\log {}'T(x_m)$	Flood discharge x_m	Serial number m	Return period $\log {}'T(x_m)$
899	1	.004	2475	57	.313
1172	2	.008	2475	58	.321
1231	3	.012	2475	59	.329
1272	4	.016	2491	60	.338
1272	5	.020	2514	61	.346
1432	6	.024	2514	62	.355
1432	7	.028	2514	63	.364
1439	8	.032	2514	64	.373
1444	9	.037	2538	65	.382
1502	10	.041	2554	66	.392
1541	11	.045	2586	67	.402
1560	12	.050	2594	68	.412
1639	13	.054	2594	69	.422
1706	14	.058	2594	70	.432
1780	15	.063	2602	71	.443
1829	16	.068	2626	72	.454
1850	17	.072	2627	73	.465
1857	18	.077	2643	74	.477
1913	19	.081	2675	75	.489
1913	20	.086	2675	76	.501
1934	21	.091	2773	77	.514
1955	22	.096	2773	78	.527
1992	23	.101	2773	79	.540
1992	24	.106	2839	80	.554
2006	25	.111	2856	81	.568
2006	26	.116	2881	82	.583
2013	27	.121	2881	83	.598
2050	28	.126	2965	84	.614
2050	29	.131	3007	85	.630
2072	30	.137	3050	86	.647
2094	31	.142	3058	87	.665
2101	32	.148	3067	88	.684
2115	33	.153	3067	89	.703
2145	34	.159	3126	90	.723
2145	35	.164	3179	91	.744
2153	36	.170	3214	92	.766

TABLE III—*Concluded*

Flood discharge x_m	Serial number m	Return period $\log {}'T(x_m)$	Flood discharge x_m	Serial number m	Return period $\log {}'T(x_m)$
2160	37	.176	3250	93	.790
2168	38	.182	3266	94	.825
2175	39	.188	3293	95	.841
2206	40	.194	3310	96	.869
2206	41	.200	3310	97	.899
2206	42	.206	3354	98	.931
2221	43	.213	3426	99	.966
2236	44	.219	3444	100	1.004
2240	45	.226	3444	101	1.045
2258	46	.232	3480	102	1.091
2281	47	.239	3606	103	1.142
2296	48	.246	3625	104	1.200
2327	49	.253	3708	105	1.267
2342	50	.260	3801	106	1.346
2358	51	.267	3810	107	1.443
2381	52	.274	3905	108	1.568
2420	53	.282	4096	109	1.744
2444	54	.289	4105	110	2.045
2452	55	.297	4390	111	
2467	56	.305			

$\Sigma x_m = 276{,}773$. $\Sigma x_m^2 = 744{,}538{,}565$.

The arithmetic mean $\bar{u}$ of distribution (22) is [4]

$$\bar{u} = u + \frac{c}{\alpha}, \tag{27}$$

where $c = 0.5772157$ is Euler's constant. The standard deviation s is

$$s = \pi/\alpha\sqrt{6}. \tag{28}$$

Therefore

$$\bar{u} = u + 0.45005s. \tag{29}$$

The reduced variable y introduced by (23) is related to the reduced variable

$$z = \frac{x - \bar{u}}{s} \tag{30}$$

by

$$z = \frac{\alpha\sqrt{6}}{\pi}(x - u) - \frac{c\sqrt{6}}{\pi}.$$

The substitution of the numerical values leads to

$$z = 0.77970y - 0.45005. \tag{30'}$$

Conversely,

$$y = 1.28255z + 0.57722. \tag{31}$$

The value (32) $v = s/\bar{u}$, the coefficient of variation, is related to the product αu. By (27) $\alpha u = \alpha\bar{u} - c$ and by (28)

$$\alpha u = \frac{\pi}{\sqrt{6}} \cdot \frac{1}{v} - c. \tag{33}$$

Therefore the numerical value of αu can also be considered as a characteristic of an observed distribution of largest values.

For the two constants we calculate for the observed distribution of largest values the two first moments

$$\bar{u} = \frac{1}{n} \sum_{m=1}^{n} x_m, \tag{34}$$

and

$$\overline{u^2} = \frac{1}{n} \sum_{m=1}^{n} x_m^2. \tag{35}$$

To get the observed standard deviation we use the Gaussian formula

$$s = \sqrt{\left(1 + \frac{1}{n-1}\right)(\overline{u^2} - \bar{u}^2)}. \tag{36}$$

According to (28) and (27)

$$\frac{1}{\alpha} = 0.7796968s, \tag{37}$$

and

$$u = \bar{u} - \frac{0.5772157}{\alpha}. \tag{38}$$

These formulas give the two constants in the distribution of largest values.

3. Flood flows interpreted as largest values. We will now apply the theory of largest values to flood flows. Let us consider the daily flow as a statistical variable, unlimited to the right. This idea is not new. The formulas proposed by Fuller [7], Hazen [15], and numerous other authors all incorporate this assumption. Gibrat [9] supposes that the daily flows vary according to Galton's distribution. Instead of postulating a specific formula for the distribution of flows we shall only suppose that it belongs to the usual exponential type, which means that condition (18) is fulfilled.

We define a flood as being the largest value of the $N = 365$ daily flows. The

flood flows are therefore the largest values of flows. This commonplace implies the distinction between floods and inundations. For each year there exists one or more floods of the same magnitude, but there might exist several different inundations or none at all. If there are several inundations in a year the greatest one will be a flood; but a flood need not to be an inundation: even a dry year has a flood. We limit ourselves to floods, assume that $N = 365$ is a large number, and represent the distribution of annual floods by the distribution (22) of largest values.

There have been objections to the concept that the daily flow is an unlimited variable. Horton [16] believes that this implies the absurd idea of unlimited floods. This opinion is shared by Slade [20], who claims that there is a definite upper limit to the magnitude of the floods for a given stream. The theory of largest values confirms only partially Horton's opinion. If we should choose distribution (20), the most probable annual flood will be limited. For this distribution, however, it might happen that the mean annual flood has no meaning. To avoid this we have chosen distribution (22), for which the mean annual flood and all the moments will be finite. A further justification of the use of (22) might be derived from the fact that Galton's distribution belongs to the exponential type. As a final argument, numerical calculations show that formula (22) gives a better fit to the observed distributions of flows.

The variable x is the annual flood flow measured in cubic meters or cubic feet per second. The mean $\bar{u}$ is the annual mean flood, whereas u is the most probable annual flood. The value s is the standard deviation of the distribution of annual floods. Finally y is called the reduced flood.

The distribution (22) possesses the properties of the observed distribution of flood flows. It is asymmetrical; rising rather quickly but falling rather slowly. The modal value is to the left of the mean (see Fig. 3).

To apply the theory of return periods let us consider the event of the highest annual discharge being greater than x. We have to replace in formula (3) the general probability $W(x)$ by the probability of flood discharges (19). The number of observations n is the number of years for which observations exist.

To use formula (3) we have to suppose that the intervals between the successive floods are all equal to one year. This assumption conforms more or less to the seasonal nature of floods.

The return period of a flood greater than x

$$T(x) = \frac{1}{1 - e^{-e^{-y}}} \tag{39}$$

is the arithmetic mean of the intervals between two years, which have a flood discharge greater than x; the discharges for the intervening years are all less than x. Therefore $T(x)$ is the mean of the number of years for which x will be surpassed once. Formula (39) gives the meaning of u from the standpoint of the return period. For $y = 0$

$$T(u) = \frac{e}{e - 1}.$$

The return period $T(u)$ of the most probable annual flood is 1.58198 years. In other words, the constant u is the flood discharge with return period

$$\log T(u) = 0.19920 \tag{40}$$

where log signifies the common logarithm. The return period of the mean annual flood is by (27) and (39) equal to 2.32762 years.

Let us now consider the relation between the flood discharge x and its return period for small and large values of x. To small values of x correspond large negative values of y and therefore return periods T approximating 1. The distribution (25) of the largest values being unlimited, the flood discharge considered as a function of log T will by (6) increase rapidly at first. To large values of x correspond large values of y and $T(x)$. If we introduce the natural logarithm, (39) gives

$$-\ln\left(1 - \frac{1}{T(x)}\right) = e^{-y}.$$

For large values of x, viz., $T(x) \geqq 10$, it is sufficiently accurate to use

$$\frac{1}{T(x)} = e^{-y},$$

so that

$$y = \ln T(x). \tag{41}$$

If the common logarithm is used,

$$\log T(x) = 0.434294\alpha(x - u). \tag{42}$$

The logarithm of the mean number of years for which the flood discharge will once be exceeded, converges towards a linear function of x. This property of the distribution of largest values was established by M. Coutagne [2]. Let us write

$$x = u + \frac{2.30258}{\alpha} \log T(x). \tag{43}$$

Then $1/\alpha$ can be considered as a measure of the increase of a flood discharge with respect to the logarithm of time.

According to the general formulas (6) and (42) the shape of the return period as a function of the flood discharge x is as follows: at the beginning i.e., for small flood discharge, the return periods are close to 1 and increase very slowly. At the end, i.e., for large flood discharges, the logarithm of the return period converges to a linear function of x.

Another form of (43) is

$$\frac{x}{u} = 1 + \frac{2.30258}{\alpha u} \log T(x). \tag{44}$$

The ratio of the flood discharge which will be exceeded in the mean once in T years to the modal annual flood converges to a linear function of the logarithm

of the return period. The constant $1/\alpha u$ of dimension zero depends, by (33), on the coefficient of variation. Its value is a characteristic of the stream. If we introduce the arithmetic mean $\bar{u}$ and the standard deviation s we obtain by (42), (27), and (28)

$$x = \bar{u} - 0.45005s + (0.77970)\,(2.30258)s \log T(x).$$

Therefore, approximately,

$$\frac{x}{\bar{u}} = 1 - \frac{9}{20}v + 1.796v \log T(x). \tag{45}$$

The right hand member of this linear equation contains only one constant, the coefficient of variation of the floods. Finally by (42) and (31)

$$\log T(x) = 0.25068 + 0.55700\frac{x - \bar{u}}{s}. \tag{46}$$

There is still another way of interpreting these asymptotic formulas. Let $T(2x)$ be the return period of the value $2x$, then by (43)

$$2x = u + \frac{\ln T(2x)}{\alpha},$$

therefore

$$2 = \frac{\alpha u + \ln T(2x)}{\alpha u + \ln T(x)},$$

and finally

$$T(2x) = T^2(x)e^{\alpha u}. \tag{47}$$

The return period of a flood of magnitude $2x$ is equal to the square of the return period of x multiplied by a factor which depends only upon the coefficient of variation.

All these asymptotic formulas are good approximations only for return periods above ten years, which means according to Table II, $y \geqq 2.25$ or according to (23), (30) and (31) $x \geqq \bar{u} + 1.3s$. The corresponding value of the flood probability is by (3) $\mathfrak{W}(x) \geqq 0.9$. The consequences of (41) can be applied to only 10% of the observations, i.e. to the large flood discharges. Their observed return periods are based on a few observations and may therefore differ considerably from the theoretical values. In spite of the above restrictions the linear formula (43) has a meaning for values of T equal to or greater than unity. We now ask: How will the most probable largest value increase with the number of observations? This number of years can again be called T. The answer to the above question requires the solution of (13′) where the distribution (25) of largest values $\mathfrak{w}(y)$ must be introduced as the initial distribution $w(x)$. From (24)

$$\frac{T-1}{e^{-e^{-y}}}\,e^{-y-e^{-y}} - 1 + e^{-y} = 0,$$

or

$$Te^{-y} = 1,$$

which is identical with (41). For $T = 1$ the most probable annual flood is of course u. Therefore the relation (41), valid for $T \geqq 1$, means: The most probable flood $u(T)$ to be reached within T years is a linear function of the logarithm of T

$$u(T) = u + \frac{2.30258 \log T}{\alpha}. \tag{41'}$$

The constant $1/\alpha$ is the slope of this straight line. The results (41–46) are related to Fuller's well-known formula [6]. This author, the first to investigate flood flows systematically, proposed a linear relation between the logarithm of the return period and the arithmetic mean of the flood discharges greater than the mth value (m taken from above). A similar empirical formula has been stated by Lane [7] and has been applied by Saville [19]. The similarities and differences between these interpolation formulas and our theory can be stated in the following way: If we start from the theory of largest values we reach these formulas as asymptotic expressions for the return period of large floods. Considered this way, our theory gives a certain justification to Fuller's hypothesis. But Fuller's and similar formulas were intended to apply to all flood discharges. Now, the distribution of the flood discharges (4) corresponding to these return periods does not fit the observations. It can be shown that these formulas involve the assumption of a simple exponential distribution $\varphi(x)$ for the flood discharges

$$\varphi(x) = \frac{1}{\bar{u} - \epsilon} e^{-(x-\epsilon)/(\bar{u}-\epsilon)}; \tag{48}$$

and the existence of a lower limit ϵ of the flood discharges given by $\epsilon = \bar{u} - s$. In Fuller's formula all flood discharges must be greater than 2/3 of the mean annual flood. The density of probability always diminishes with increasing magnitude of the flood. This neglects the ascending branch (about one third) of the distribution of floods (see Fig. 3) and is incompatible with the observed facts. We therefore prefer our formula which takes account of the total variation, but we do not minimize the importance of Fuller's work which has led to much valuable research.

Formula (39) gives the theoretical return periods $T(x)$ as a function of the reduced flood discharge y, and holds for the entire range of observations. The general numerical values are given in Table II, cols. 1 and 3. For a given stream, the return period of a flood discharge greater than x depends by (23) upon the two constants α and u. If these values have been calculated by (37) and (38) the theoretical flood discharge x corresponding to $T(x)$ is obtained by the linear transformation

$$x = u + y/\alpha. \tag{49}$$

The asymptotic formula (42) suggests the coordination of the flood discharges x and the logarithm of the return periods.

4. Rhône and Mississippi Rivers. We think that our system of formulas is simple, logically consistent and free of artificial assumptions. Now it remains to be shown that the arithmetic involved is simple and that the results fit the observations. For the Rhône we shall analyze the observed cumulative frequency, the distribution, and the return periods. For the Mississippi River we shall limit ourselves to the return periods.

For each year we choose the maximum of the daily discharges (we do not use momentary peaks). The 111 values x_m for the Rhône 1826–1936 published by Coutagne [3] and arranged in order of increasing magnitude are given in Table III (col. 1). The supposition that the intervals between consecutive floods are all equal to one year is not always true. Only 77 of the 111 floods occurred between October and March, whereas 34 were scattered throughout the year. But the

TABLE IV

Calculation of constants

Stream observation station	Rhône Lyon (France) 1826–1936	Mississippi River Vicksburg (Miss.) 1890–1939
Number of observations n	111	50
Annual mean flood $\bar{u}$	2,493.5	1,355.6
Mean squared flood $\overline{u^2}$	6,707,555.0	1,951,828.8
Standard deviation s	703.1	341.3
Constant $1/\alpha$	548.2	266.1
Most probable annual flood u	2,177.0	1,201.9

differences in the lengths of the intervals compensate each other. The second column of Table III contains the serial number m. According to (9) we calculate for the mth observed flood discharge x_m, taken in ascending magnitude, the logarithm of the observed return period $\log n/(n - m)$ (col. 3), where $n = 111$ and $m = 1, 2, \cdots, 110$, and obtain the exceedance intervals. The other observed curve, the recurrence interval, is obtained by (10) through the coordination of x_{m+1} and $\log n/(n - m)$. Both curves are plotted in Fig. 1. The recurrence and exceedance intervals differ for the large flood discharges. The observed flood discharges arranged in increasing magnitude are plotted in the cumulative histogram, Fig. 2.

To compare these observations with our theory, we calculate the two constants $1/\alpha$ and u according to the formulas (34)–(38). The values Σx_m and Σx_m^2 are given at the end of Table III. Division by $n = 111$ gives the mean flood $\bar{u}$ and the mean squared flood $\overline{u^2}$ (Table IV). The Gaussian correction being $1 + 1/110$ we obtain from formula (36) the standard deviation s (Table IV)

TABLE V

Observed and theoretical distributions of flood discharges

Rhône

Reduced variable y	Variable x	Midpoints $x+\frac{\Delta x}{2}$	Observed distribution $111\Delta'\mathfrak{W}(x)$	Theoretical distribution $111\Delta\mathfrak{W}(x)$	Cumulative frequency $111\mathfrak{W}(x)$
−2.75	670				
−2.50		807	1		0.00
−2.25	944			0.01	0.01
−2.00		1081	1	0.34	0.07
−1.75	1218			1.19	0.35
−1.50		1355	7	3.03	1.26
−1.25	1492			6.07	3.38
−1.00		1629	5	9.98	7.33
−0.75	1766			14.02	13.36
−0.50		1903	13	17.38	21.35
−0.25	2040			19.49	30.74
0.00		2177	21	20.21	40.84
0.25	2314			19.68	50.95
0.50		2451	19	18.26	60.52
0.75	2588			16.31	69.21
1.00		2725	14	14.14	76.83
1.25	2862			11.97	83.35
1.50		2999	9	9.94	88.80
1.75	3136			8.15	93.29
2.00		3273	8	6.61	96.95
2.25	3410			5.30	99.90
2.50		3547	6	4.23	102.25
2.75	3686			3.45	104.13
3.00		3822	4	2.65	105.70
3.25	3959			2.00	106.78
3.50		4096	2	1.64	107.70
3.75	4233			1.28	108.42
4.00		4370	1	1.01	108.98
4.25	4507			0.79	109.43
4.50		4644	0	0.61	109.77
4.75	4781			0.48	110.04
5.00		4918		0.38	110.25
5.25	5055			0.30	110.42
5.50		5192		0.23	110.55
5.75	5329			0.18	110.65
6.00		5466		0.27	110.73
			111	111.00	

and finally from (37) and (38) the constant $1/\alpha$ and the most probable annual flood u. From the numerical values in Table IV the linear transformation (49) for the Rhône is

$$x = 2177.03 + 548.19y.$$

TABLE VI

Observed return periods

Mississippi River, Vicksburg, (Miss.) (1890–1939)

Flood discharge x_m	Serial number m	Return period $\log' T(x_m)$	Flood discharge x_m	Serial number m	Return period $\log' T(x_m)$
760	1	0.0088	1357	26	.3188
866	2	.0178	1457	27	.3273
870	3	.0269	1397	28	.3566
912	4	.0362	1397	29	.3768
923	5	.0458	1402	30	.3980
945	6	.0555	1406	31	.4202
990	7	.0655	1410	32	.4437
994	8	.0758	1410	33	.4686
1018	9	.0862	1426	34	.4949
1021	10	.0969	1453	35	.5229
1043	11	.1079	1475	36	.5529
1057	12	.1192	1480	37	.5851
1060	13	.1308	1516	38	.6198
1073	14	.1427	1516	39	.6576
1185	15	.1549	1536	40	.6990
1190	16	.1675	1578	41	.7448
1194	17	.1805	1681	42	.7959
1212	18	.1939	1721	43	.8539
1230	19	.2076	1813	44	.9208
1260	20	.2219	1822	45	1.0000
1285	21	.2366	1893	46	1.0969
1305	22	.2518	1893	47	1.2219
1332	23	.2676	2040	48	1.3980
1342	24	.2840	2056	49	1.6990
1353	25	.3011	2334	50	

$\Sigma x_m = 67{,}780.\quad \Sigma x_m^2 = 97{,}591{,}440.$

This leads to the determination of the theoretical flood discharges. The theoretical return periods $\log T(x)$ are given in Table II, col. 3 as a function of the reduced variable y and of x (col. 4). The discharges x obtained by letting y take on the values -2.75 to 6.00 in the linear transformation, are given in

Table V, cols. 2 and 3 and plotted in Fig. 1. The distances Δx used in the calculations of the theoretical discharges are $1/4\alpha = 137.05$.

Along the abscissa are plotted the logarithm of the return periods and the return periods in years; along the ordinate are plotted the corresponding flood discharges and the modal annual flood u. The straight line from the point $(u, 0)$ to the asymptote gives the most probable flood as a function of time. The theoretical curve corresponds quite closely with the general course of the observations. For small floods the theoretical return periods are practically iden-

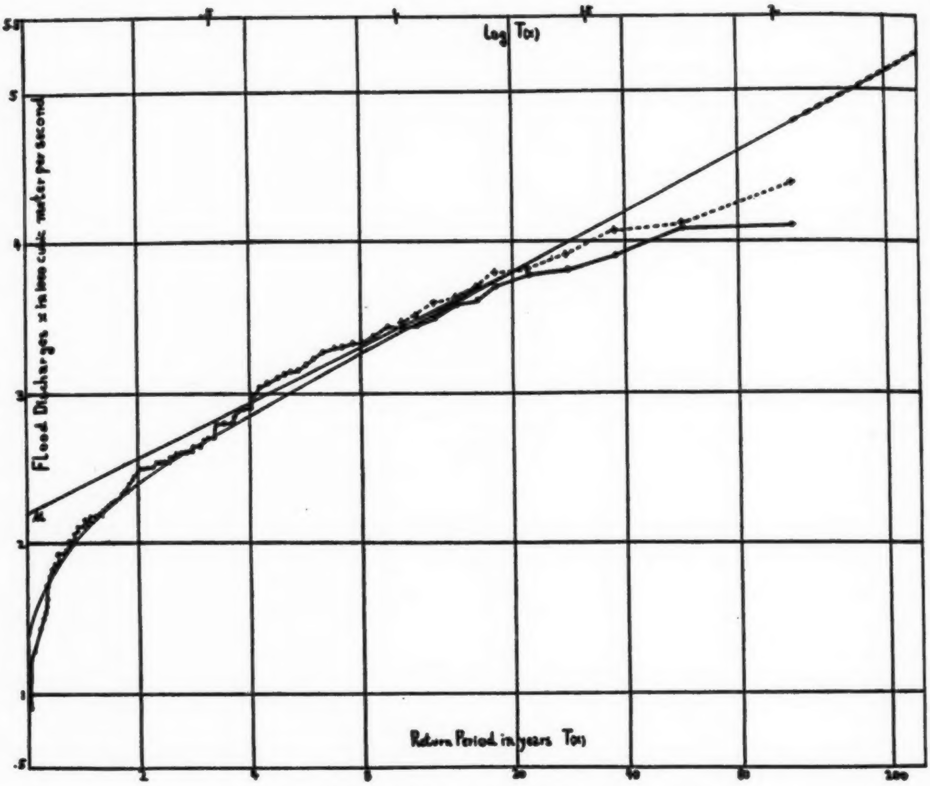

FIG. 1. RHÔNE AT LYON (FRANCE) 1826–1936

Observations Table III: Recurrence intervals, + – – +; Exceedance intervals, •——•; Return periods, ——; Theory Table II, cols. 3 and 4: Extrapolation, — —.

tical with the observed values. But for the very large floods the theoretical curve surpassed both the exceedance and recurrence intervals.

The observed cumulative histogram is shown in Fig. 2. We calculate from Table II, col. 2, the frequencies $111\mathfrak{W}(x)$ (Table V, col. 6). These theoretical values $(x, 111\mathfrak{W}(x))$ are also plotted in Fig. 2. The agreement between theory and observations is very good.

For the comparison of the observed and theoretical distributions of the flood discharges we use what might be called the natural classification. For the

observations, the length of the class intervals and the beginning of the first class interval are arbitrary. In order to obtain the observed distribution of the flood discharges, it is natural to use the theoretical class intervals set forth in Table V, col. 2. The data of the third column can be interpreted as the midpoints of the class intervals given in col. 2. The frequencies for these class intervals are ob-

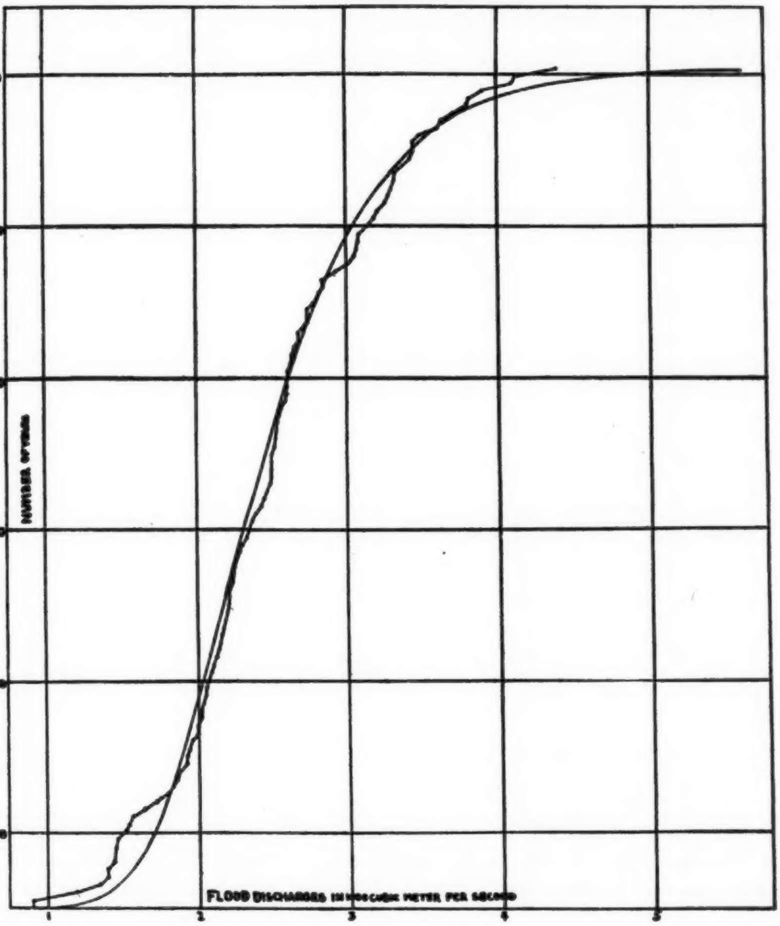

FIG. 2. CUMULATIVE FREQUENCY OF THE FLOOD DISCHARGES. RHÔNE, LYON (FRANCE) 1826–1936

Observations Table III cols. 1 and 2, •—•; Theory Table V cols. 2, 3 and 6, /

tained from Table III, and are given in Table V, col. 4. The observed distribution is shown in Fig. 3. To obtain the corresponding theoretical distribution we calculate from Table V, col. 6, the difference between two cumulative frequencies disjoined by one, i.e., we pair consecutively the first and third, the second and fourth items and so on. This theoretical distribution given in col. 5 and the observed distribution are based on class intervals of the same length. Fig. 3

shows that the theoretical distribution $\Delta\mathfrak{W}(x)$ of the largest values agrees in a satisfactory way with the observed distribution $\Delta'\mathfrak{W}(x)$ of the flood discharges. Table VI, col. 1, gives the corrected[2] flood discharges x_m, measured in units of 1000 cubic feet per second, for the Mississippi River at Vicksburg (1890–1939), ($n = 50$), arranged according to increasing magnitude; col. 2 gives the serial number m. We calculate the logarithm of the observed return periods log $n/(n - m)$, (col. 3). The observations (x_m, log $'T(x_m)$) and (x_{m+1}, log $'T(x_m)$) are plotted in Fig. 4. The constants obtained by formulas (34)–(38) are shown

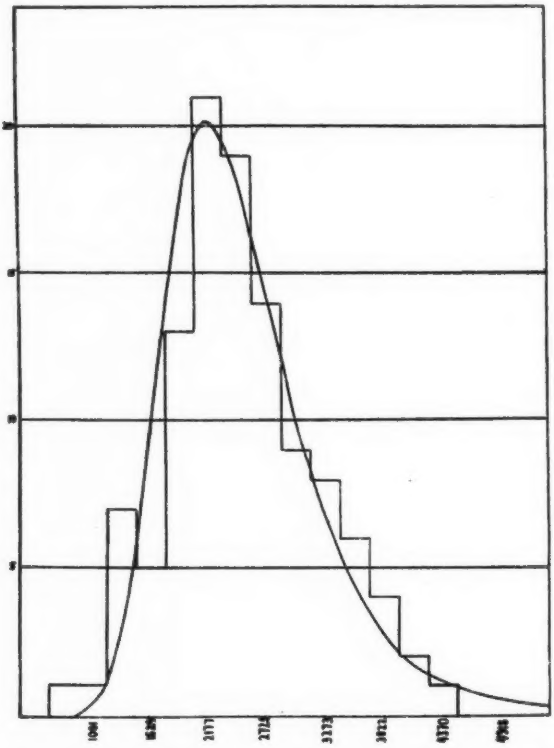

FIG. 3. DISTRIBUTION OF THE FLOOD DISCHARGES. RHÔNE, LYON (FRANCE) 1826–1936 Observations Table V cols. 2, 3 and 4, ⊓; Theory Table V cols. 2, 3 and 5,

in Table IV. By (49) the theoretical floods x corresponding to the return periods $T(x)$ presented in Table II, col. 3, are

$$x = 1201.98 + 266.14y.$$

These floods are given in Table II, col. 5. The class interval used is

$$1/4\alpha = 66.5.$$

[2] These data have been put at my disposal through the courtesy of Mr. A. E. Brandt of the U. S. Department of Agriculture.

The theoretical curve $(x, \log T(x))$, plotted in Fig. 4, agrees in a very satisfactory way with the observations. For the large floods the theoretical return periods are between the exceedance and recurrence intervals.

The calculations of the theoretical return periods for other streams, e.g. the Columbia, Connecticut, Cumberland, Rhine, and Tennessee Rivers, for which reliable observations exist for more than 60 years, also show a good agreement with the observations. The goodness of fit diminishes for streams for which the number of observations is smaller and for which the data are not very reliable.

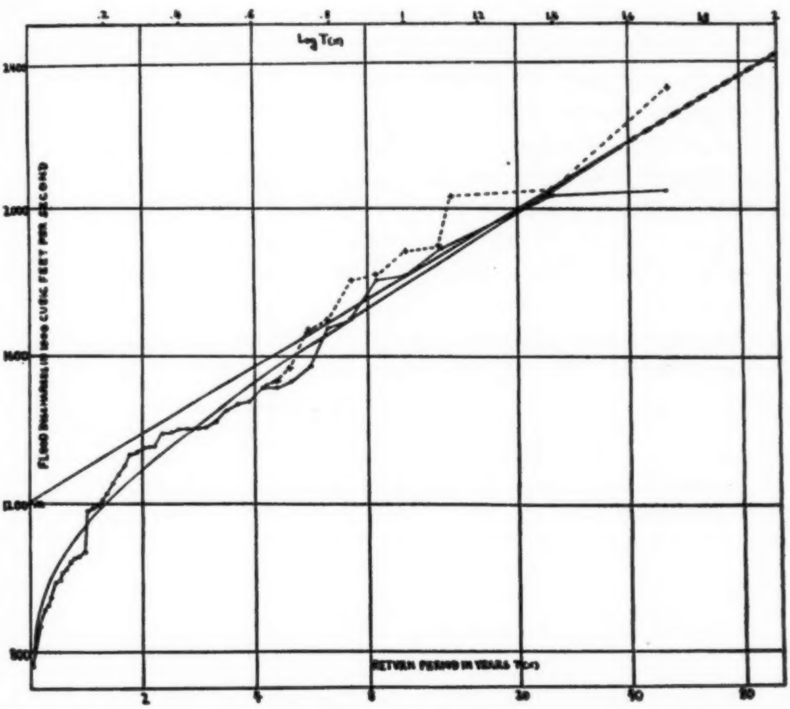

FIG. 4. MISSISSIPPI RIVER AT VICKSBURG, (MISS.) 1890–1939

Observations Table VI: Recurrence intervals, + — — +; Exceedance intervals, •——•; Return periods, ———; Theory Table II, cols. 3 and 5; Extrapolation, — —.

5. Summary and conclusions. In order to apply any theory we have to suppose that the data are homogeneous, i.e. that no systematical change of climate and no important change in the basin have occurred within the observation period and that no such changes will take place in the period for which extrapolations are made. It is only under these obvious conditions that forecasts can be made.

The theoretical return period $T(x)$, the mean number of years between two annual flood discharges greater than or equal to x, is a statistical function such as the distribution $w(x)$ or the probabilities $W(x)$ and $P(x)$. There are two

sets of observed values corresponding to the theoretical set. The exceedance interval $'T(x_m)$ formula (9), and the recurrence interval $''T(x_m)$ formula (9'); x_m being the mth flood discharge, where m is counted from below. As any theory must include both notions, no separate theory for exceedance or recurrence intervals is possible.

The return period $T(x)$ of a flood discharge x is found by formula (39). For large values of x the flood discharge converges toward a linear function (42) of the logarithm of the return period. This is the scientific basis of Fuller's empirical formula. The two constants of our formula u and $1/\alpha$, are, respectively, the most probable annual flood discharge and a multiple of the standard deviation (28). Their values depend upon the drainage basin and known geological and meteorological factors. It is beyond our present task to consider the influence of these factors. Our method can be summarized by the following rules:

1) For each year find the maximum daily discharge x_m (do not use momentary peaks) and arrange these n data in increasing magnitudes.

2) Calculate for each discharge x_m ($m = 1, 2, \cdots, n - 1$), the values $\log {}'T(x_m) = \log n - \log (n - m)$ and plot the curves x_m, $\log n/(n - m)$, and x_{m+1}, $\log n/(n - m)$. These are the observed exceedance and recurrence intervals.

3) Calculate the annual mean flood $\bar{u}$ and the annual mean squared flood $\overline{u^2}$; determine according to (36)–(38) the standard deviation

$$s = \sqrt{\left(1 + \frac{1}{n - 1}\right)(\overline{u^2} - \bar{u}^2)}$$

and the two constants

$$1/\alpha = 0.77970s,$$

$$u = \bar{u} - \frac{0.57722}{\alpha}.$$

4) The theoretical flood discharges x corresponding to the logarithm of the return period $T(x)$ given in Table II, col. 3, are obtained by the linear transformation

$$x = u + y/\alpha$$

where y is taken from Table II, col. 1. Plot x as a function of $\log T(x)$. For large values of x and for extrapolation it is sufficient to use the linear asymptote obtained graphically.

The linear part of the theoretical curve (x, $\log T$) permits of two interpretations: First, T is the theoretical return period of a flood greater than or equal to x; second, x is the most probable flood to be reached within T years. The second interpretation holds for the straight line through the point (u, 0).

The figures show a close agreement between observed and theoretical values.

The observed curvature of the return periods is brought out by the theoretical graph.

The agreement between theory and observation is excellent for floods which correspond to reduced values of $y \leqq 3$. For the two or three extreme floods, the return periods are based on a few observations and, consequently, the agreement is not very good. No theory can be verified by two or three observations. Generally speaking, the theory fits the observations as closely as could be expected for such a complicated phenomenon.

In order to make a further test of our results, we need a numerical measure for the weights to be given to the theoretical points. Therefore, for a given probability we must find the corresponding theoretical limits for the observed return periods. The theory of positional values will give these control curves. Since it was the purpose of this article to develop and make clear the basic method, we have refrained from introducing this subject.

It is our claim that the calculus of probabilities and especially the theory of largest values, is an efficient tool for the solution of certain hydrological problems.

REFERENCES

[1] G. E. Becker and C. E. van Orstrand, *Hyperbolic Functions*, Smithsonian Mathematical Tables, Washington, 1931.

[2] A. Coutagne, "Etude statistique des débits de crue," *Revue Générale de l'Hydraulique* Paris (1937).

[3] A. Coutagne, "Etude statistique et analytique des crues du Rhône à Lyon," *Comptes Rendus du Congrès pour l'Utilisation des Eaux*, Lyon, (1938).

[4] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the smallest and the largest member of a sample," *Proc. Camb. Phil. Soc.*, Vol. 24 (1928).

[5] M. Fréchet, "Sur la loi de probabilité de l'écart maximum," *Annales Soc. Polon. Math.*, Vol. 6, (1927).

[6] Weston E. Fuller, "Flood flows," *Trans. Am. Soc. Civil Eng.*, Vol. 77 (1914).

[7] Weston E. Fuller, E. Lane and others, "Discussion on flood flow characteristics," *Trans. Am. Soc. Civil Eng.*, Vol. 89 (1926).

[8] John C. Geyer, "New curve fitting method for analysis of flood-records," *Trans. Am. Geophy. Union*, Part II (1940), pp. 660–668.

[9] Robert Gibrat, "Aménagement hydro-électrique des cours d'eau," "Statistique mathématique et calcul des probabilités, "*Revue Générale de l'Electricité*, Vol. 32, No. 15, 16, Paris (1932).

[10] Eugene L. Grant, "The probability-viewpoint in hydrology," *Trans. Am. Geophy. Union*, Part 1 (1940), pp. 7–12.

[11] E. J. Gumbel, "Les valeurs extrêmes des distributions statistiques," *Annales de l'Institut Henri Poincare*, Vol. 4 (1935), p. 115.

[12] E. J. Gumbel, "La plus grande valeur," *Aktuarske Vedy*, Vol. 5, No. 2, p. 83, No. 3, p. 133, No. 4, p. 146, Prague (1935–36).

[13] E. J. Gumbel, *La Durée Extrême de la Vie Humaine, Actualités Scientifiques et Industrielles*, Hermann et Cie, Paris, 1937.

[14] E. J. Gumbel, "Les intervalles extrêmes entre les émissions radioactives," *Jour. de Phys.*, Serie 7, Vol. 8, No. 8, No. 11 (1937).

[15] Allen Hazen, *Flood Flows, A Study of Frequencies and Magnitudes*, John Wiley and Sons, Inc., New York, 1930.

[16] ROBERT E. HORTON, "Hydrologic conditions as affecting the results of the application of methods of frequency analysis to flood records," *Geological Survey Water-Supply Paper* 771, Washington (1936).

[17] CLARENCE S. JARVIS, "Floods in the United States, Magnitude and Frequency," *Geological Survey Water Supply Paper* 771, Washington (1936).

[18] R. VON MISES, "La distribution de la plus grande de *n* valeurs," *Revue Math. de l'Union Interbalkanique*, Vol. 1, Athens (1936).

[19] T. SAVILLE, "A study of methods of estimating flood flows applied to the Tennessee River," *Publications from College of Engineering*, Nr. 6, New York (1935–36).

[20] J. J. SLADE, "The reliability of statistical methods in the determination of flood frequencies," *Geological Survey Water-Supply Paper* 771, Washington (1936).

ON THE FOUNDATIONS OF PROBABILITY AND STATISTICS[1]

BY R. VON MISES

Harvard University

1. Introduction. The theory of probability and statistics which I have been upholding for more than twenty years originates in the conception that the only aim of such a theory is to give a description of certain observable phenomena, the so called mass phenomena and repetitive events, like games of chance or some specified attributes occurring in a large population. Describing means here, in the first place, to find out the relations which exist between sequences of events connected in some way, e.g. a sequence of single games and the sequence composed of sets of those games or between a sequence of direct observations and the so called inverse probability within the same field of observations. The theory is a mathematical one, like the mathematical theory of electricity, based on experience, but operating by means of mathematical processes, particularly the methods of analysis of real variables and theory of sets.

We all know very well that in colloquial language the term probability or probable is very often used in cases which have nothing to do with mass phenomena or repetitive events. But I decline positively to apply the mathematical theory to questions like this: What is the probability that Napoleon was a historical person rather than a solar myth? This question deals with an isolated fact which in no way can be considered as an element in a sequence of uniform repeated observations. We are all familiar with the fact that, e.g. the word energy is often used in every day language in a sense which does not conform to the notion of energy as adopted in mathematical physics. This does not impair the value of the precise definition of energy used in physics and on the other hand this definition is not intended to cover the entire field of daily application of the term energy.

We discard likewise the scholastic point of view displayed in a sentence of this kind: "... that both in its meaning and in the laws which it obeys, probability derives directly from intuition and is prior to objective experience." This sentence is quoted from a mathematical paper printed in a mathematical journal of 1940. The same author continues calling probability a metaphysical problem and speaking of the difficulties "which must in the nature of things always be encountered when an attempt is made to give a mathematical or physical solution to a metaphysical problem." In my opinion the calculus of probability has nothing to do with metaphysics, at any rate not more than geometry or mechanics has.

[1] Address delivered on September 11, 1940 at a meeting of the Institute of Mathematical Statistics in Hanover, N. H.

On the other hand we claim that our theory, which serves to describe observable facts, satisfies all reasonable requirements of logical consistency and is free from contradictions and obscurities of any kind. I am now going to outline the essential ideas of the theory as developed by me since 1919 and I shall have to refer as to the proof of its consistency to the recent work of A. H. Copeland, of J. Herzberg and of A. Wald. Then I will give some examples of application in order to show how the theory works and how it applies to actual problems in statistics.

2. The notion of kollektiv. The basic notion upon which the theory is established is the concept of *kollektiv.* We consider an infinite sequence of experiments or observations every one of which supplies a definite result in the form of a number (or a group of numbers in the case of a kollektiv of more than one dimension). We shall designate briefly by X the sequence of results x_1, x_2, x_3, $\cdots$. In tossing a die we get for X an endless repetition of the integers one to six, $x = 1, 2, \cdots 6$. If we are interested in death probability, we observe a large group of healthy 40 year old men and mark a one for each individual surviving his 41st aniversary and a zero for each man who dies before, so that the sequence x_1, x_2, x_3, $\cdots$ consists of zeros and ones. In a certain sense the kollektiv corresponds to what is called a *population* in practical statistics. Experience shows that in such sequences the relative frequency of the different results (one to six in the first of our examples, one and zero in the second) varies only slightly, if the number of experiments is large enough. We are therefore prompted to assume that in the kollektiv, i.e. in the theoretical model of the empirical sequences or populations, each frequency has a *limiting value,* if the number of elements increases endlessly. This limiting value of frequency is called, under certain conditions which I shall explain later, the "probability of the attribute in question within the kollektiv involved." The set of all limiting frequencies within one kollektiv is called its *distribution.*

Let me insist on the fact that in no case is a probability value attached to a single event by itself, but only to an event as much as it is the element of a well defined sequence. It happens often that one and the same fact can be considered as an element of different kollektivs. It may then be that different probability values can be ascribed to the same event. I shall give a striking example of this, which we encounter in the field of actual statistical problems, at the end of this lecture.

The objection has been made: Since all empirical sequences are obviously finite sequences, why then assume infinite kollektivs? Our answer is that any straight line we encounter in reality has finite length, but geometry is based on the notion of infinite straight lines and uses e.g. the notion of parallels which has no sense, if we restrict ourselves to segments of finite lengths. Another objection, often repeated, reads that there is a contradiction between the existence of a frequency limit and the so called Bernoulli theorem which states that sequences of any length showing a frequency say $\frac{1}{2}$ can also occur in cases for

which the probability equals $\frac{1}{2}$. But it has been proved, in a rigorous way excluding any doubt, that the two statements are compatible, even by explicit construction of infinite sequences fulfilling both conditions. I would even claim that the real meaning of the Bernoulli theorem is inaccessible to any probability theory that does not start with the frequency definition of probability.

Now we are in the position to explain how our probability theory works. This sequence of zeros and ones

$$(X) \qquad 1\ 0\ 1\,|\,0\ 0\ 1\,|\,1\ 0\ 0\,|\,0\ 1\ 1\,|\,1\ 1\ 0\,|\,0\ 1\ 1\,|\,0\ 1\ 0\,|\,1\ 1\ 1\ \cdots$$

may represent the outcomes of a game of chance. The ones show gains, the zeros losses for one of the two players. If we separate the terms of X into groups of three digits and replace each group by a single one or zero according to the majority of terms within the group, we get a new sequence

$$(X') \qquad 1\ 0\ 0\ 1\ 1\ 1\ 0\ 1\ \cdots$$

which represents the gains and losses in sets of three games. Our task is now to compute the distribution, i.e. the limiting frequencies of zeros and ones in this new sequence X', assuming the two frequencies in X are known. A sequence can formally be considered as a unique number like a decimal fraction with an infinite number of digits. Then the transition from X to X' can be called a *transformation of a number* $X' = T(X)$. As our sequences have to fulfill certain conditions Copeland calls the sequences X, X' admissible numbers. What I just quoted was of course a very special example of a transformation of a number. But we have to emphasize that all problems dealt with in probability theory, without any exception, have this unique form: The distribution or the limiting frequencies in certain sequences are given, other sequences are derived from the given ones by certain operations, and the distributions in these derived sequences have to be computed. In other words: *Probability theory is the study of transformations of admissible numbers, particularly the study of the change of distributions implied by such transformations.*

We know four and only four simple, i.e. irreducible transformations or *four fundamental operations.* They are called selection, mixing, partitioning and combination. By combining these basic processes we can settle all problems in probability theory. The formal, mathematical difficulties in carrying out the computation of the new distributions may become very serious in certain cases, particularly if we have to apply an infinite number of transformations (asymptotic problems). But, in the clearly defined framework of this theory no space is left for any metaphysical speculations, for ideas about sufficient reason or insufficient reason, for notions like degree of evidence or for a special kind of probability logic and so on. And further no modification is needed for handling usual statistical problems: Terms like inverse probability, likelihood, confidence degrees, etc. are justified and admitted only as far as they are capable of being reduced to the basic notion of kollektiv and distribution within a kollektiv. I will give some more details to this point later. Meanwhile let me turn to a

general question which, in a certain way, is the crucial point in establishing the new probability theory.

3. Place selections and randomness. It is obvious that we have to restrict still further the notion of kollektiv or the field of sequences which can be considered as the objects of a probability investigation. The successive outcomes of a game of chance differ very clearly from any regular sequence as defined by a simple arithmetical law, e.g. the regularly alternating sequence 0 1 0 1 0 1 0 1 $\cdots$. A typical property which singles out the irregular or random sequences and which has to be reproduced in every probability theory is that, if p is the probability of encountering a one in the sequence, then p^2 is the probability of two ones following each other immediately. Any probability theory has to introduce an axiom which enables us to deduce this theorem and others of a similar type. The question is only how to find a sufficiently general and consistent form for it. The procedure I have chosen consists in using a special kind of transformation of a sequence, which I call a *place selection*.

A place selection is defined by an infinite set of functions $s_n(x_1, x_2, \cdots x_{n-1})$ where $x_1, x_2, x_3, \cdots$ are the digits of an admissible number or a kollektiv and s_n has one of the two values zero or one. Here $s_n = 1$ means that the nth digit of the sequence is retained, $s_n = 0$ means that it is discarded. The decision about retaining or discarding the nth elements depends as you see, only on the preceding values $x_1, x_2, \cdots x_{n-1}$, but not on x_n or the following digits. Example of a place selection:

$$\begin{aligned} s_n = 1, &\text{ if } x_{n-1} = 0 \text{ for prime numbers } n,\\ &\text{ if } x_{n-1} = 1 \text{ for } n \text{ not prime},\\ s_1 = 1, &\text{ and } s_n = 0 \text{ in all other cases.}\end{aligned}$$

Experience shows that, if we apply such a place selection to the sequence X of outcomes of a game of chance, we get a new, selected sequence $S(X)$ in which the frequencies of gains and losses are about the same as in X. This fact or the practical *impossibility of a gambling system* suggests the adoption of the following procedure in handling transformations of admissible numbers.

First, if within a certain investigation the transformation applied to X is a place selection, we assume that the distribution in $X' = S(X)$ is the same as in X: distr $S(X) =$ distr X. Second, if a general transformation T is applied to X, say $X' = T(X)$, then we examine whether the existence of a place selection S that changes the distribution in X' (so as to have distr $S(X') \neq$ distr X') implies the existence of a place selection S_1 that would affect the distribution in X (so as to give distr $S_1(X) =$ distr X). If this is the case, we say that X' is a kollektiv, provided that the original sequence X was considered to be a kollektiv. Take e.g. for X the sequence resulting from tossing a die endlessly, and call $p_1, p_2, \cdots p_6$ the limiting frequencies of the six possible outcomes $1, 2, \cdots 6$. The transformation T may consist in replacing every 1 in the sequence X by a

2, every 3 by a 4, and every 5 by a 6. The new sequence consists of only three different kinds of elements 2, 4, 6 and therefore its distribution includes only three values p_2', p_4', p_6' where evidently $p_2' = p_1 + p_2$ etc. Here it is almost obvious that if a place selection applied to X' changes the value of p_2', the same selection if applied to X must change either p_1 or p_2. So, if the original sequence X was considered as a kollektiv, X' has to be admitted too.

Now the question arises whether this procedure is in itself consistent or whether it can lead to contradictions. We were concerned up to now with kollektivs the elements of which belong to a finite set of distinct numbers $e_1, e_2, \cdots e_k$ and the distributions of which are therefore defined by k non-negative values $p_1, p_2, \cdots p_k$ with the sum 1. In this case it was pointed out by Wald and by Copeland that, if an arbitrary distribution and an arbitrary countable set Σ of place selections are given, there exists a continuum of sequences every one of which has the given distribution, which is not affected by any place selection belonging to Σ. Now it may be supposed that in a concrete problem a sequence X' is derived from a sequence X by a finite number of fundamental operations involving a finite set Σ' of place selections. Another finite set Σ'' may consist of selections employed in establishing that certain sequences used in the derivation of X' are "combinable" ones. Finally an arbitrary countable set Σ of selections S may be assumed. According to our procedure we have shown that to any place selection S which affects the distribution in X' corresponds a certain S_1 which, when applied to X, changes the distribution of X. All these S_1 corresponding to the elements S of Σ form a countable set Σ_1. Now the set Σ_2 including Σ', Σ'', Σ_1 and also including all products of two of its own elements is a countable set too. What we use in computing the distribution of X' is only the fact that the given sequence X is unaffected by the selections that are elements of Σ_2. It follows from the above quoted results that we can substitute for X a numerically specified sequence and carry out all operations upon this specified sequence. So it is proved that no contradiction can arise in computing the final probability according to our conception.

I cannot enter here into a discussion of the more complicated case where the range within which the elements of a kollektiv vary, is an infinite one, either a countable set or a continuum. All principal problems connected with establishing the notion of kollektiv can be settled satisfactorily, at any rate, by considering those general forms of sequences as limiting cases of kollektivs with a finite set of attributes.

4. Example: Set-of-games problem. I want to present now a simple, but instructive example to show how the theory works and what task a mathematical foundation of the calculus of probability has to achieve. Let us recall the two sequences X and X' composed of zeros and ones of which we spoke above. The first represented the outcomes of a sequence of single games, the second the outcomes of triple sets of those games. If X is considered as a kollektiv with

given probabilities p and q for one and zero, it is easy to deduce the corresponding values p' and q' for X' and to show that X' is a kollektiv too. We begin by carrying out three selections which single out from the original sequence x_1, x_2, $x_3 \cdots$ first, the elements x_1, x_4, x_7, $\cdots$ second, the elements x_2, x_5, x_8, $\ldots$ and third, the elements x_3, x_6, x_9, $\cdots$. It can be shown by means of certain further place selections that these three kollektivs which we call X_1, X_2, X_3 are combinable. That means that combining the corresponding elements of the three sequences like $x_1x_2x_3$, $x_4x_5x_6$, $x_7x_8x_9$, $\cdots$ leads to a new three dimensional kollektiv X_0 in which each permutation of three digits 0 and 1, has a probability equal to the corresponding product of p- and q-factors. For instance the probability of encountering the group 111 is p^3 and for the group 110 it is p^2q. Now we operate a mixing upon X_0 by collecting all permutations with two or three ones. We find in a well known way the sum $p^3 + 3p^2q$ for the probability p' of ones in the sequence X'. So far the result is very well known and can be reached—in my opinion, in a very incomplete and unsatisfactory way—also by the classical methods.

But what I want to discuss here is a slightly modified question. If the sequence X means gains and losses for single games and if the arrangement for sets of three games is made as indicated before, then in a real play the gains and losses of sets are counted in a different way. For, if the first two games of a set are both won or lost by the same player, the fate of the set is decided and there is no sense to play the third game. So the loss of the second set in our example will already be recognized after the fifth game and the actual sixth game will be considered as the first game of the third set. In this way the original sequence X decomposed into groups of two or three games

$$(X) \qquad 1\ 0\ 1 \mid 0\ 0 \mid 1\ 1 \mid 0\ 0 \mid 0\ 1\ 1 \mid 1\ 1 \mid 0\ 0 \mid 1\ 1 \mid 0\ 1\ 0 \mid 1\ 1 \mid \cdots$$

leads to a new sequence X''

$$(X'') \qquad 1\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ \cdots$$

which is obviously different from X'. Everyone familiar with the usual handling of the probability concept will say that in X'' the probabilities of zeros and ones must be the same as in X'. But a mathematical foundation of theory of probability, if it deserves this name, has to clear up the question: From what principles or particular assumptions and by what inferences may we deduce the equality of the limiting frequencies in X' and X''?

There is no difficulty in solving this problem from the point of view of the frequency theory. We have only to apply somewhat different place selections instead of the above used which lead to the kollektivs X_1, X_2, X_3. I showed elsewhere how the general set-of-games problem can be satisfactorily treated in this way. Here I want to stress only that the problem as a whole is completely inaccessible by any of the other known approaches to probability theory. The classical point of view which starts with the notion of equally likely cases and rests upon a rather vague idea of the relationship between probability and

sequences of events does not even allow the formulation of the problem. In the so called modernized classical theory, as proposed by Fréchet, probabilities are defined as "physical magnitudes of which frequencies are measures." Fréchet would say that the frequencies both in X' and in X'' are measures of the same quantity. But why? We face here obviously a mathematical question which cannot be settled by referring to physical facts. It is clear that the equality of the distributions in the two sequences X' and X'' is due to the randomness or irregularity of the original sequence X. No theory which does not take in account the randomness, which avoids referring to this essential property of the sequences dealt with in probability problems, can contribute anything toward the solution of our question.

I have to make some special remarks about the so-called measure theory of probability.[2]

5. Probability as measure. Up to now we have been concerned only with the simplest type of kollektivs, namely, with those sequences the elements of which belong to a finite set of numbers so as to have a distribution consisting of a finite number of finite probabilities with the sum 1. It may be true that all practical problems, in a certain sense, fall into this range. For, the single result of an observation is always an integer, the number of smallest units accessible to the actual method of measuring. Nevertheless in many cases it is much more useful to adopt the point of view that the possible outcomes of an experiment belong to a more general set of numbers, e.g. to a continuous segment or any infinite variety. If we include the case of kollektivs of more than one dimension, we have to consider a point set in a k-dimensional space (where even k may be infinite) as the label set or attribute set of the kollektiv. In order to define the probability in this case we have to choose a subset A of the label set and to count among the first n elements the number n_A of those elements the attributes of which fall into A. Then the quotient $n_A : n$ is the frequency, and its limiting value for n infinite will be called the probability of the attribute falling into A within the given kollektiv.

It was rightly stressed by many authors that in the case of an infinite label set some additional restrictions must be introduced. In particular A. Kolmogoroff set up a complete system of such restrictions. We cannot ask for the existence of the limiting frequency in any arbitrary subset A. It will be sufficient to assume that the limit exists for a certain Körper or a certain additive family of subsets. If it exists for two mutually exclusive subsets A and B, the limit corresponding to $A + B$ will be, by virtue of the original definition, the sum of the limits connected with A and B. We can now insert a further axiom involving the complete additivity of the limiting values. So we arrive at the statement

[2] What I call measure theory here is essentially that proposed by Kolmogoroff in his pamphlet of 1933. As to the new theory developed by Doob in his following paper (where instead of the label space the space of all logically possible sequences is used in establishing the measures) see my comment on page 215.

that probability is the measure of a set. All axioms of Kolmogoroff can be accepted within the framework of our theory as a part of it, but in no way as a substitute for the foregoing definition of probability.

Occasionally the expression probability as measure theory is used in a different sense. One tries to base the whole theory on the special notion of a set of measure zero. One of the basic assumptions in my theory is that in the sequence of results we obtain in tossing a so called correct die the frequency, say of the point 6, has a certain limiting value which equals 1/6. A different conception consists in stating that anything can happen in the long run with a correct die, even that an uninterrupted sequence of six's or an alternating sequence of two's and four's or so on may appear. Only all these events which do not lead to the limiting frequency 1/6 form, together as a whole, a set of events of measure zero. Instead of my assumption: the limiting value is 1/6 we should have to state: It is almost certain that a limit exists and equals 1/6. Nothing can be said against such an alluring assumption from an empirical standpoint, since actual experience extends in no case to an infinite range of observations. The only question is whether the asumption is compatible with a complete and consistent theory. I cannot see how this may be achieved. Before saying that a set has measure zero we have to introduce a measure system which can be done in innumerable ways. If e.g. we denote the outcome six by a one and all other outcomes 1 to 5 by zero, we get as the result of the game with a die an infinite sequence of zeros and ones. It has been shown by Borel that according to a common measure system the set of all 0, 1 sequences which do not have the limiting frequency $\frac{1}{2}$ has the measure zero. In this way it turns out to be almost certain that the limiting frequency of the outcome six in the case of a correct die is $\frac{1}{2}$. Other values for the limit can be obtained by a similar inference. It is a correct but misleading idea that the measure zero is unaffected by a regular (continuous) transformation of the assumed measure system, since in our field of problems different measures which are not obtained from one another by a regular transformation have equal rights. So, saying that a certain set has the measure zero makes in our case no more sense than to state that an unknown length equals 3 without indicating the employed unit.

In recapitulating this paragraph I may say: First, the axioms of Kolmogoroff are concerned with the distribution function within one kollektiv and are *supplementary to my theory, not a substitute for it.* Second, using the notion of measure zero in an absolute way without reference to the arbitrarily assumed measure system, *leads to essential inconsistencies.*

6. Statistical estimation. Let me now turn to the last point, the application of probability theory to one of the most widely discussed questions in today's statistical research: the so-called estimation problem. Many strongly divergent opinions are facing each other here. I think that the probability theory based on the notion of kollektiv is best able to settle the dispute and to clear up the difficulties which arose in the controversies of different writers.

We may, without loss of generality, restrict ourselves to the simplest case of a single statistical variable x and a single parameter ϑ, where x of course may be the arithmetical mean of n observed values. Here (and likewise in the case of more variables and more parameters) we have to distinguish carefully among four different kollektivs which are simultaneously involved in the problem. The range within which both x and ϑ vary will be assumed to be a continuous interval so that all distributions will be given by probability densities.

The first kollektiv we deal with is a one-dimensional one where the probability of x falling into the interval x, $x + dx$ depends on x and on a parameter ϑ. If

$$p(x \mid \vartheta) \tag{1}$$

denotes the corresponding density and the limits A, B within which x possibly falls depend on ϑ too, we have

$$\int_{A(\vartheta)}^{B(\vartheta)} p(x \mid \vartheta)\, dx = 1 \qquad \text{for each } \vartheta. \tag{1'}$$

In order to fix the ideas we may imagine that the first kollektiv consists in drawing a number x out of an urn and that ϑ characterizes the contents of the urn. Asking for an estimate of ϑ implies the assumption that different possible urns are at our reach every one of which can be used for drawing the x. The ϑ values for the different urns fall into a certain interval C, D. It is usual to suppose that the urns are picked out at random so as to give another one-dimensional kollektiv with the independent variable ϑ. Let $p_0(\vartheta)\, d\vartheta$ be the probability of picking an urn with the characteristic value falling into the interval ϑ, $\vartheta + d\vartheta$. This density

$$p_0(\vartheta) \tag{2}$$

is often called the *prior* or *a priori* probability of ϑ. As the range within which ϑ varies is confined by the constants C and D, we have obviously

$$\int_C^D p_0(\vartheta)\, d\vartheta = 1. \tag{2'}$$

Now from these two one-dimensional kollektivs with the variables x in the first, ϑ in the second, we deduce by combination (multiplication) a two-dimensional kollektiv with the density function

$$P(\vartheta, x) = p_0(\vartheta) \cdot p(x \mid \vartheta). \tag{3}$$

The individual experiment which forms the element of this third kollektiv consists of picking at random an urn and drawing afterwards from this urn. Both x and ϑ are now independent variables (attributes of the kollektiv) and it is easy to see that it follows from (1) and (2)

$$\int_C^D \int_{A(\vartheta)}^{B(\vartheta)} P(\vartheta, x)\, dx\, d\vartheta = \int_C^D p_0(\vartheta)\, d\vartheta \int_{A(\vartheta)}^{B(\vartheta)} p(x \mid \vartheta)\, dx = 1. \tag{3'}$$

We will return later to this two-dimensional kollektiv. Let us, first, derive from it, by applying the operation of partitioning (Teilung), our fourth and last kollektiv which is one-dimensional again. Partitioning means that we drop from the sequence of experiments which form the third kollektiv all those for which the x-value falls outside a certain interval x, $x + dx$; and that in this way we consider a partial sequence of experiments with only the one variable ϑ. The distribution of ϑ-values within this sequence with quasi-constant x is given, according to the well known rule of division or rule of Bayes (a rule which can be proved mathematically) by[3]

$$p_1(\vartheta \mid x) = \frac{P(\vartheta, x)}{\int_C^D P(\vartheta, x)\, d\vartheta} = c(x)\, p_0(\vartheta)\, p(x \mid \vartheta). \tag{4}$$

It follows immediately that

$$\int_C^D p_1(\vartheta \mid x)\, d\vartheta = 1. \tag{4'}$$

This function p_1 of ϑ depending on the parameter x is generally called the *posterior* or *a posteriori* probability of ϑ.

If $p_1(\vartheta \mid x)$ can be computed according to the formula (4), every question concerning the "presumable" value of ϑ as drawn from the outcome x of an experiment is completely answered. We can find indeed, by integration the probability which corresponds to any part of the interval C, D of ϑ and so the estimation problem is definitely solved. But the trouble is that in most cases of practical application nothing or almost nothing is known about the prior probability $p_0(\vartheta)$ which appears as a factor in the expression of p_1. Hence arises the new question: *What can we say about the ϑ-values without having any information about its prior probability?* This is the estimation problem as it is generally conceived today.

The first successful approach to the answering of this question was made by Gauss. If we do not know p_1, we know however, except for a constant factor, the quotient p_1/p_0, posterior probability to prior probability which equals $cp(x \mid \vartheta)$. The maximum of this quotient must be greater than one, since the average values of both p_0 and p_1 are the same. So the maximum means the point of the greatest increase produced by the observed experimental value of x upon the probability of ϑ. It seems reasonable to assume the ϑ-value for which the ratio p_1/p_0 reaches its maximum as an estimate for ϑ: It is the value upon which the greatest emphasis is conferred by the observation. This idea, originally proposed by Gauss in his theory of errors, has been later developed chiefly by R. A. Fisher, and is known today as the maximum likelihood method. Calling the ratio p_1/p_0 likelihood seems indeed an adequate nomenclature.

---

[3] For brevity Bayes' rule is employed in the text as in the case of a discontinuous distribution. The correct procedure in the case of a continuous x would require that we first use finite intervals and then pass to the limit.

The method of estimation used most frequently today is not the maximum likelihood method, but the so called confidence interval method, inaugurated by R. A. Fisher and now successfully extended and applied by J. Neyman. This method uses the third of the above mentioned kollektivs instead of the fourth, i.e. the two-dimensional probability $P(\vartheta, x)$. At first sight it seems hopeless to use this function which includes the unknown prior probability $p_0(\vartheta)$ as a factor. But it turns out as Neyman has shown[4] (and this is the decisive idea of the confidence interval method) that we can indicate in the x, ϑ-plane special regions for which the probability $\iint P(\vartheta, x)\,dx\,d\vartheta$ is independent of $p_0(\vartheta)$. In fact, if we point out for every ϑ such an interval x_1, x_2 as to have

$$\int_{x_1(\vartheta)}^{x_2(\vartheta)} p(x \mid \vartheta)\,dx = \alpha, \qquad 0 < \alpha < 1, \tag{5}$$

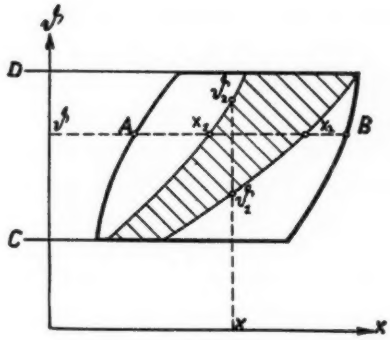

Fig. 1

it follows immediately from (2) and (5) for the region covered by these intervals

$$\int_C^D \int_{x_1(\vartheta)}^{x_2(\vartheta)} P(\vartheta, x)\,dx\,d\vartheta = \int_C^D p_0(\vartheta)\,d\vartheta \int_{x_1(\vartheta)}^{x_2(\vartheta)} p(x \mid \vartheta)\,dx = \alpha. \tag{6}$$

For given α the intervals can be chosen in different ways. If we choose $x_1 = A$ for $\vartheta = C$ and $x_2 = B$ for $\vartheta = D$, we get a strip or belt, as shown in Fig. 1 which supplies for every given x a smallest value ϑ_1 and a greatest value ϑ_2. The definition of our third kollektiv leads to the conclusion: *If we predict each time a certain x is observed that ϑ lies between the corresponding ϑ_1 and ϑ_2, then the probability is α that we are right, whatever the prior probability may be.*[5] It is

[4] J. Neyman, *Roy. Stat. Soc. Jour.*, Vol. 97 (1934), pp. 590–92.

[5] After my lecture Dr. A. Wald called my attention to Neyman's suggestion; namely that this statement can be generalized by admitting that the infinite sequence of ϑ-values which results from picking out successively the urns for drawing a number x, does not fulfill the conditions of a kollektiv. So, instead of the terms "whatever the prior probability may be" we can say "whatever the method of picking out the urns may be." In fact, let us consider the case where ϑ can assume only a finite number of values $\vartheta_1, \vartheta_2, \cdots \vartheta_k$. Among the n first trials let n_κ be the number of cases where $\vartheta = \vartheta_\kappa$ and $n'_\kappa \leq n_\kappa$ the number of cases where $\vartheta = \vartheta_\kappa$ and x falls into the interval $x_1(\vartheta_\kappa)$, $x_2(\vartheta_\kappa)$. The relative

understood that in this argument both x and ϑ are variables the values of which may change from one trial to the next. I cannot agree with the statement, which is often made, that x only is a variable and ϑ a constant or that we are only interested in one specified value of ϑ. In no way is it possible, in the framework of the confidence limits method, to avoid the idea of a so-called superpopulation, i.e. the existence of a manifold of urns every one of which forms a kollektiv.[6] Thus no contradiction and no antagonism exists between this method and the Bayes formula. Only a different kollektiv, a two-dimensional instead of a one-dimensional, is here considered.

I have no time to enter here in a discussion of the very interesting developments of Neyman's theory which are intended to supply additional conditions in order to determine the arbitrary choice of the x-intervals in a unique way. May I only mention that what is called in Neyman's theory the probability of a second type error in testing the hypothesis $\vartheta = \vartheta_0$ is given by the expression

$$(7) \qquad \int_C^D \int_{x_1(\vartheta_0)}^{x_2(\vartheta_0)} P(\vartheta, x)\, dx\, d\vartheta = \int_C^D p_0(\vartheta)\, d\vartheta \int_{x_1(\vartheta_0)}^{x_2(\vartheta_0)} p(x \mid \vartheta)\, dx.$$

If we want to determine the confidence belt or the intervals x_1, x_2 in such a way as to minimize this expression independently of the function $p_0(\vartheta)$, we obtain Neyman's maximum power condition

$$(8) \qquad \int_{x_1(\vartheta_0)}^{x_2(\vartheta_0)} p(x \mid \vartheta)\, dx \equiv F(\vartheta, \vartheta_0) = \text{min. for each pair } \vartheta, \vartheta_0.$$

This condition, it is well known, cannot be fulfilled under general assumptions for $p(x \mid \vartheta)$. Moreover the above-mentioned boundary conditions $x_1(C) = A(C)$ and $x_2(D) = B(D)$ (or similar ones in other cases) have to be considered too. If they are not satisfied, the statement which can be made with probability α would include the prediction that certain x-values are impossible. Except for this case the above formulated theorem is equally valid for every region determined according to (5).

It is clear that if the original distribution is given by a regular, slightly varying function $p(x \mid \vartheta)$, the confidence limits method cannot give very substantial results. Let us take e.g. for $p(x \mid \vartheta)$ the uniform distribution

$$(9) \qquad p(x \mid \vartheta) = 1/\vartheta \text{ for } 0 \leqq x \leqq \vartheta, \qquad 0 \leqq \vartheta \leqq 1.$$

---

frequency of correct predictions is then $(n_1' + n_2' + \cdots n_k'): n$ where n equals $n_1 + n_2 + \cdots n_k$. If n tends to infinity, at least one part of the n_κ must become infinite. For those the limit of $n_\kappa' : n_\kappa$ tends to α according (5) while the other terms (with finite n_κ and n_κ') have no influence. So the limiting value of the frequency $(n_1' + n_2' + \cdots n_k'): n$ equals in any event α. This generalization does not apply, if we ask for the probability of a second type error of the hypothesis $\vartheta = \vartheta_0$. Here the existence of the prior probability p_0 is essential.

[6] According to the generalization supplied by Neyman's point of view (*Phil. Trans. Roy. Soc.*, Vol. A-236 (1937), pp. 333–380) which is discussed in footnote 5, the superpopulation does not necessarily satisfy the conditions of a kollektiv.

We have here $A = 0$, $B = \vartheta$, $C = 0$, $D = 1$ and the domain in which x and ϑ vary is the 45° right triangle shown in Fig. 2. Whatever $p_0(\vartheta)$ may be, the integral of $p(\vartheta, x) = p_0(\vartheta)\cdot p(x \mid \vartheta)$ over this domain is 1 and if we omit the part of the triangle on the left of the straight line $x = (1 - \alpha)\vartheta$, the integral over the remaining part is α. For $\alpha = 0.90$, a statement which can be made with a probability of 90% reads: The value of ϑ lies between x and $10x$. On the other hand we know from the very beginning with 100% certainty that ϑ lies between x and 1, so that for $x \geqq 0.1$ the statement is futile. (If one chooses as confidence belt the part on the left of the straight line $x = \alpha\vartheta$, the statement would run: ϑ lies between 1.1 x and 1 and values of x greater than 0.9 are impossible.) If we apply in this case the Bayes formula, we find that the outcome depends to the highest extent on what is known about the prior probability $p_0(\vartheta)$.

In most cases however which present themselves in practical statistics the original density function $p(x \mid \vartheta)$ has a different character from that assumed in

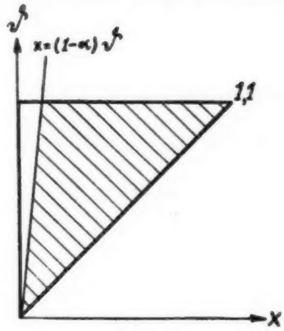

FIG. 2

(9). It depends generally on an integer n and the distribution is concentrated more and more when n increases. (We may define here concentration as standard deviation tending towards zero. The integer n means in general the number of basic experiments). We have e.g. in the so-called Bayes problem where x is the arithmetical mean of n observations the asymptotic expression for p:

$$p(x \mid \vartheta) \sim \sqrt{\frac{n}{2\pi\vartheta(1-\vartheta)}}\, e^{-\frac{1}{2}n(x-\vartheta)^2/\vartheta(1-\vartheta)} \tag{10}$$

$$0 \leqq \vartheta \leqq 1, \qquad 0 \leqq x \leqq 1.$$

If we denote by Φ the probability integral

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2}\, du, \tag{11}$$

the x-intervals corresponding to a given probability value α are defined by

$$x_1 = \vartheta - \xi, \qquad x_2 = \vartheta + \xi \qquad \text{where } \Phi\left(\xi \sqrt{\frac{n}{2\vartheta(1-\vartheta)}}\right) = \alpha. \tag{12}$$

If n has a large value, the ξ's are very small and we get a narrow belt along the straight line $x = \vartheta$ as shown in Fig. 3 for $\alpha = 0.90$ and n about 100. The prediction which can be made with the probability α reads approximately

$$x - \eta \leqq \vartheta \leqq x + \eta \qquad \text{where } \Phi\left(\eta \sqrt{\frac{n}{2x(1-x)}}\right) = \alpha. \tag{13}$$

On the other hand it is well known that in this case the Bayes formula supplies a posterior probability $p_1(\vartheta \mid x)$ which turns out to be more and more independent of the prior probability $p_0(\vartheta)$ when n increases. It has been shown that the asymptotic expression for $p_1(\vartheta \mid x)$ whatever $p_0(\vartheta)$ may be, is

$$p_1(\vartheta \mid x) \sim \sqrt{\frac{n}{2\pi x(1-x)}}\, e^{-\frac{1}{2}n(\vartheta - x)^2/x(1-x)}. \tag{14}$$

It follows that, on the basis of the Bayes formula, we can predict for every single value of x with the probability α that ϑ lies between the above given

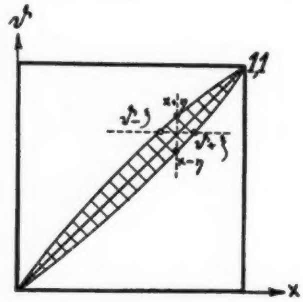

FIG. 3

limits (13). This is more than the confidence limits method supplies, but the result is subjected to the restriction that $p_0(\vartheta)$ is a continuous function. However, for large values of n (generally this means for large numbers of basic experiments) the outcomes of both methods are essentially the same.

Let me recapitulate in three brief sentences the essential results we have found in the problem of estimation.

1. There is no contradiction of any kind between the Bayes formula and the confidence limits method and no difference at all in the underlying probability concept. In both methods the idea of a sort of "super-population" is used. Only two different kollektivs are considered in both cases.

2. If the original distribution has a regular, slightly varying density function $p(x \mid \vartheta)$, the Bayes method gives a complete answer when the prior probability is known and no answer when it is unknown. The confidence limits method gives in both cases a definite solution; it lies in the nature of things that the solution cannot be very substantial if $p(x, \vartheta)$ is only slightly varying.

3. If the original distribution $p(x \mid \vartheta)$ depends on a further parameter n and becomes concentrated more and more with increasing n, both approaches give, for large n, asymptotically about the same results.

It is not intended by these remarks to impair the value of the confidence limits method which both from theoretical and from practical point of view deserves our attention. But the rather inconceivably aggressive attitude towards the Bayes' theory as displayed by a number of statisticians, which, however, does not include J. Neyman, turns out to be completely unfounded.

PROBABILITY AS MEASURE

By J. L. Doob

University of Illinois

The following pages outline a treatment of probability suitable for statisticians and for mathematicians working in that field. No attempt will be made to develop a theory of probability which does not use numbers for probabilities. The theory will be developed in such a way that the classical proofs of probability theorems will need no change, although the reasoning used may have a sounder mathematical basis. It will be seen that this mathematical basis is highly technical, but that, as applied to simple problems, it becomes the set-up used by every statistician. The formal and empirical aspects of probability will be kept carefully separate. In this way, we hope to avoid the airy flights of fancy which distinguish many probability discussions and which are irrelevant to the problems actually encountered by either mathematician or statistician.

We shall identify as Problem I the problem of setting up a formal calculus to deal with (probability) numbers. Within this discipline, once set up, the only problems will be mathematical. The concepts involved will be ordinary mathematical ones, constantly used in other fields. The words "probability," "independent," etc. will be given mathematical meanings, where they are used.

We shall identify as Problem II the problem of finding a translation of the results of the formal calculus which makes them relevant to empirical practice. Using this translation, experiments may suggest new mathematical theorems. If so, the theorems must be stated in mathematical language, and their validity will be independent of the experiments which suggested them. (Of course, if a theorem, after translation into practical language, contradicts experience, the contradiction will mean that the probability calculus, or the translation, is inappropriate.)

The classical probability investigators did not separate Problems I and II carefully, thinking of probability numbers as numbers corresponding to events or to hypothetical truths, and always referring the numbers back to their physical counterparts. The measure approach to the probability calculus has put this approach into abstract form, and separated out the empirical elements, thus removing all aspects of Problem II. We shall explain this approach first in a simplified set-up, that which will be made to correspond (Problem II) to a repeated experiment in which the results of the nth trial can be any integer x_n between 1 and N (inclusive), in which the experiments are independent of each other, and performed under the same conditions. (The set-up will be applicable, for example, to the repeated throwing of a die.)

The measure approach treats this experiment as follows. Let $\omega: (x_1, x_2, \cdots)$ be any sequence of integers between 1 and N, inclusive. We consider ω as a point in an infinite dimensional space Ω. (Each point ω may be considered as a logically possible sequence of results of the given experiment, and this fact will guide us in solving Problem II.) A measure function is defined on certain sets of points of Ω as follows. Let $p_1, \cdots, p_N$ be any numbers satisfying the conditions

$$p_j \geqq 0, \qquad j \geqq 1, \qquad p_1 + \cdots + p_N = 1.$$

(How these numbers are chosen in any particular problem will be explained below. The method of choice is irrelevant to the mathematics, but is involved in the solution of Problem II.) The set of all sequences beginning with $x_1 = \alpha$ is given measure p_α. More generally, the measure of the set of all sequences beginning with $x_1 = \alpha_1, \cdots, x_n = \alpha_n$, is defined as $p_{\alpha_1} \cdot p_{\alpha_2} \cdots p_{\alpha_n}$. In this way, as can be shown,[1] a completely additive measure function is determined on certain point sets of Ω, on a field $\mathfrak{F}$ of sets so large that all the usual Lebesgue measure and integration theory is applicable. This means that there is a collection $\mathfrak{F}$ of sets of points of Ω such that if $S_1, S_2, \cdots$ are finitely or infinitely many sets in the collection, their sum $\sum_1^\infty S_n$, their intersection $\prod_1^\infty S_n$, and their complements are also in the collection. Each set S in $\mathfrak{F}$ has a definite measure $P(S)$, $0 \leqq P(S) \leqq 1$, and if $S_1, S_2, \cdots$ are finitely or infinitely many disjunct sets in $\mathfrak{F}$,

$$P(S_1 + S_2 + \cdots) = P(S_1) + P(S_2) + \cdots.$$

Problem II, the translation problem, is solved as follows. Each relevant event is made to correspond to a point set of Ω. A relevant event is a physical concept—defined by imposing some set C of conditions on the results of the experiments. The corresponding Ω-set is the set of sequences $(x_1, x_2, \cdots)$ satisfying the same set C of conditions, imposed on the x_j. Thus the set of all sequences beginning with $x_1 = \alpha_1$, $x_2 = \alpha_2$, is made to correspond to the event: *the result of the first experiment is α_1, of the second is α_2*. As is to be expected, the mathematical picture goes further than the real one. The "event" *1 occurs infinitely often in a sequence of trials* has only conceptual significance, physically, but the corresponding point set of Ω: the set of all sequences $(x_1, x_2, \cdots)$ containing infinitely many 1's, is a perfectly definite point set whose measure can be calculated in terms of $p_1, \cdots, p_N$. (In fact it is easily seen that this measure is 1 or 0, according as $p_1 > 0$ or $p_1 = 0$.) By "the probability of an event" we shall mean the measure of the corresponding Ω-set. As this measure has been defined, the probability that the nth trial results in a number j is p_j, and the probability that one trial results in j, and another in k, is $p_j \cdot p_k$.

[1] Cf. A. Kolmogoroff, *Ergebnisse der Mathematik*, Vol. 2, No. 3, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, where the most complete treatment of the approach to the probability calculus from the standpoint of measure is given.

The justification of the above correspondence between events and Ω-sets is that certain mathematical theorems can be proved, filling out a picture on the mathematical side which seems to be an approximation to reality, or rather an abstraction of reality, close enough to the real picture to be helpful in prescribing practical rules of statistical procedure. The following two theorems are important ones, from this point of view. These two theorems depend in no way on observed facts. They are stated and proved in the customary language of modern analysis.

THEOREM A: Let j_n be the number of the first n coordinates of the point $\omega:(x_1, x_2, \cdots)$ which are equal to j, where j is some integer $(1 \leqq j \leqq N)$ which will be kept fixed throughout the discussion. Then $0 \leqq j_n \leqq n$, and j_n varies from point to point on $\Omega: j_n = j_n(\omega)$ is a function of ω, that is of the sequence $(x_1, x_2, \cdots)$. When $n \to \infty$, j_n/n has not a unique limit independent of the sequence $(x_1, x_2, \cdots)$ under consideration. In fact if ω is the point $(k, k, \cdots)$, $j_n(\omega) = 0$ for all n, unless $j = k$; if ω is the point $(j, j, \cdots)$, $j_n(\omega) = n$ for all n. It is simple to give examples of sequences $\omega:(x_1, x_2, \cdots)$ for which $j_n(\omega)$ oscillates without approaching a limit, as $n \to \infty$. But Theorem A (usually called the strong law of large numbers) states that there is a set of sequences, i.e. an ω-set S, of measure 0, such that

$$\lim_{n \to \infty} \frac{j_n(\omega)}{n} = p_j, \tag{1}$$

unless ω is in S. In other words the sequences for which (1) is not true are exceptional in the sense of measure theory. If a new choice $\{p_i'\}$ of p_i's is made, then if $p_j' \neq p_j$, the new exceptional set includes all the sequences which were not exceptional before, since the limit in (1) becomes p_j'. Thus S depends essentially on p_j. Theorem A is a generalization of Bernoulli's classical theorem which states in our language that the measure of the set of sequences $\omega:(x_1, x_2, \cdots)$ for which

$$|j_n(\omega)/n - p_j| > \epsilon$$

approaches 0, as $n \to \infty$, for any positive ϵ. Theorem A is stronger because it states that there is actual convergence, whereas Bernoulli's theorem only concludes that there is a kind of convergence on the average.

Theorem A corresponds to certain observed facts, relating to the clustering of "success ratios," giving rise to empirical numbers $\bar{p}_j$. If the statistician wishes to apply his calculus to a given experiment (Problem II), he sets $p_i = \bar{p}_j$. There has been frequent discussion of the problem of determining the $\bar{p}_j$. This discussion of the $\bar{p}_j$ is sometimes held on so high a plane that the innocent bystander may wonder to what purpose such abstract philosophic concepts could possibly be put—besides that of stimulating further discussion on a still higher plane. The principle purpose of this paper is to discuss Problem I, but a few words on Problem II might not be out of place here. Almost everyone who is going to use probability numbers, the $\bar{p}_j$, *for other than conversational purposes,*

derives them in the same way. There is a judicious mixture of experiments with reason founded on theory and experience. Thus if a coin is tossed by an experimenter who has examined the coin, and found that it had heads on one side but not on both, that it seemed balanced, and that (as a confirming check) tossing a hundred times gave around 50 heads, the experimenter would use $\frac{1}{2}$ as the probability of obtaining heads in his further reasoning. Of course there is no logic compelling this. The experimenter may have been fooled. A coin far out of balance may turn up 50 heads in 100 throws. But man must act, and the above procedure has been found useful, which is all that is desired. In many experiments, less reliance can be placed on a preliminary physical examination of the experimental conditions, and more must be placed on the actual working out of the experiment, as in the analysis of machine products. In that case, the actual results must be examined with great care, before attempting to use the above mathematical set-up. It sometimes may even be possible to change the experimental conditions to make the mathematics applicable.[2] In all cases, such mathematical theorems as Theorem A and the following Theorem B give the basis for applying the formal apparatus to practice. Indeed, the criterion of application includes the verification of special cases of the practical versions of Theorems A and B.

THEOREM B: Let $f_n(x_1, \cdots, x_{n-1})$ $(n > 1)$ be any function of the indicated variables, except that we suppose f_n only takes on the values 0, 1. Let $\omega:(x_1, x_2, \cdots)$ be a given point of Ω. Let n' be the number of the first n integers i such that $f_i(x_1, \cdots, x_{i-1}) = 1$, and let j'_n be the number of the first n integers i such that $f_i(x_1, \cdots, x_{i-1}) = 1$, and $x_i = j$. Then j'_n, n' are functions of $\omega:(x_1, x_2, \cdots)$. If $f_1 \equiv f_2 \equiv \cdots \equiv 1$, $j'_n = j_n$, $n' = n$, where j_n is as defined above. Suppose that there is an Ω-set S_0 of measure 0 such that $n' \to \infty$, as $n \to \infty$, unless $\omega \epsilon S$. Theorem B states that there is then an Ω-set S' of measure 0, such that if $\omega:(x_1, x_2, \cdots)$ is not in S',

$$\lim_{n\to\infty} \frac{j'_n(\omega)}{n'} = p_j. \tag{1'}$$

(The set S' will depend on the given functions $f_1, f_2, \cdots$ and on the p_i, but is fixed, once these have been chosen.) This mathematical theorem corresponds to certain observed facts (usually summarized by stating that no (successful) system of play is possible). In fact, it states, in the language of practice, that rejecting certain trials, using as a criterion of acceptance or rejection the results of preceding trials, rejecting the ith trial if $f_i(x_1, \cdots, x_{i-1}) = 0$, does not affect the outcome of a game of chance, or, more precisely, does not affect the validity of the physical fact corresponding to Theorem A. If $f_1 \equiv f_2 \equiv \cdots \equiv 1$, (1′) becomes (1). The hypothesis that $n' \to \infty$ as $n \to \infty$ unless $\omega \epsilon S_0$ is made to insure that infinitely many trials will be accepted. As an example of the

[2] Cf. W. A. Shewhart, *Statistical Method from the Viewpoint of Quality Control*, Washington, 1939.

possible variety in the definition of the f_i, we might define f_i as 1 if $x_{i-1} = N$, and $f_i = 0$ otherwise, so trials are accepted only if the previous trial resulted in the number N. Or much more complicated systems can easily be devised in which the criterion of acceptance of the nth trial depends on a varying number of the results of preceding trials. This theorem gives a mathematical counterpart to the physical idea of the mutual independence of repeated trials.

To summarize, mathematically (Problem I) the study has been reduced to that of the measure properties of Ω. This can be considered independently of any physical correspondence. The physical correspondence (Problem II) makes any event $\mathfrak{E}$ correspond to a point set E of Ω, the "probability of $\mathfrak{E}$" becomes the measure of E. Thus "the probability that the result of the first experiment is 3" becomes the measure of the set of sequences $(x_1, x_2, \cdots)$ beginning with $x_1 = 3$. *We have given no sharp definition of probability as a physical concept.* If the above mathematical set-up, after translation, using some set of p_j's, seems to fit a given physical set-up, any event will be said to have as its probability, the measure of the corresponding Ω-set. We have attempted to give no intrinsic a priori definition of the probability of an event: such a definition is quite unnecessary for our purposes. All that was required was a basis for prescribing the usual statistical procedures, and we have described such a basis.

In the above example, there would have been no new difficulty introduced if the x_n were not restricted to integral values, but allowed to take on any numerical values. The general point $\omega:(x_1, x_2, \cdots)$ of Ω would now be any sequence of real numbers. Instead of choosing the numbers $p_1, \cdots, p_N$ we choose a "distribution function" $F(x)$, a monotone function with the following properties:

$$\lim_{x\to-\infty} F(x) = 0, \qquad \lim_{x\to+\infty} F(x) = 1, \qquad F(x-0) = F(x).$$

Measure on Ω is defined as follows. The set of all sequences beginning with x_1 such that $a \leqq x_1 < b$ is given measure $F(b) - F(a)$. (The number $F(b)$ is called "the probability that $x_1 < b$.") More generally, the measure of the set of all sequences $(x_1, x_2, \cdots)$ beginning with $x_1, \cdots, x_n$, such that $a_j \leqq x_j < b_j$, $j = 1, \cdots, n$ is defined as $\prod_j [F(b_j) - F(a_j)]$. Thus if $F(x)$ defines a simple rectangular distribution: $F(x) = 0$ for $x < 0$, $F(x) = x$ for $0 \leqq x \leqq 1$, $F(x) = 1$ for $x > 1$, Ω-measure becomes (infinite dimensional) volume in the (infinite dimensional) unit cube. The correspondence (Problem II) between events and point sets of Ω is defined just as before. Sometimes it may be useful, in considering experiments giving rise to pairs of numbers, to let each x_n be a pair of numbers so that Ω becomes a sequence of points of a plane instead of a sequence of points of a line. In all cases there are mathematical theorems true of the resulting Ω which guide us (Problem II) in deciding just how the Ω-measure is to be defined, that is, how $F(x)$ is to be defined, in dealing with a given practical problem. But the essential point is this. Once Ω-measure has been defined, no changes or further hypotheses are possible or necessary. All

relevant probability questions are answerable. Thus consider a question of the following type: if the experiments are grouped in some way,[3] with what probability will the groups have some given regularity property?[4] The question singles out a set E of sequences of Ω and asks: what is the measure of E? The problem may or may not be difficult mathematically,[5] depending on the grouping, but the original definition of measure on Ω needs no enlargement to answer it.

Technically, the mathematics has become the mathematics of a special type of measure defined on a space of infinitely many dimensions. If, however there is an integer ν such that only at most ν experiments are to be considered, we need only consider the ν-dimensional space of points $(x_1, \cdots, x_\nu)$, defining measure in this space in the same way as on Ω. Thus if x_n has the rectangular distribution defined above, the measure in $(x_1, \cdots, x_\nu)$-space becomes ordinary ν-dimensional volume in the unit cube. Perhaps the most common measure a statistician considers is that in which the measure of an $(x_1, \cdots, x_\nu)$-set E becomes "the probability that the point $(x_1, \cdots, x_\nu)$ representing an independent sample of ν from a normal distribution of mean 0 and variance σ^2" will lie in E:

$$P\{E\} = \sigma^{-\nu}(2\pi)^{-\frac{1}{2}\nu} \int \underset{E}{\cdots} \int e^{-\frac{1}{2}(x_1^2+\cdots+x_\nu^2)/\sigma^2}\, dx_1 \cdots dx_\nu. \tag{2}$$

This example makes it obvious that the statistician is always doing measure theory, even though he may not state that fact explicitly. If the number of experiments has no upper bound conceptually—mathematically when the number of dimensions ν may increase without limit, as in Theorems A, B, it is much more convenient to use the space Ω, in terms of which experiments with varying numbers of trials can be considered simultaneously. The classical proofs of probability theorems, such as Bernoulli's theorem (the law of large numbers) are perfectly correct. If the "probability of an event" is interpreted as the measure of a set, these proofs do not even need verbal changes. There can be no question of the need for any axiomatic development beyond that necessary for measure theory, and the probability calculus can lead to no contradiction, unless the theory of measure is faulty.

It is customary for probability theorists to stop their discussions when the present stage is reached, so that the beginnings of a formal calculus have been constructed to deal with a repetition of independent experiments, conducted

---

[3] A grouping is necessary, for example, when two players are playing a game in which two out of three wins in the trials win a game. The trials are then grouped into successive groups of two or three, depending on how they come out.

[4] Continuing the preceding note, the question might be: will the ratio (games won by player α)/(games played) approach a limit with probability 1, that is, for all of the original sequences $\{x_n\}$ except possibly some forming a set of measure 0?

[5] The answer to the question of the preceding notes is simple. If p is the probability that player α wins a trial, the ratio in question approaches $p^3 + 3p^2(1 - p)$, the probability that α wins a game, with probability 1.

under the same conditions. Perhaps this is because of the following widely held syllogism: probability is something dealing with random events; random events are events having no influence on each other; therefore . . . . Unfortunately mathematicians and statisticians must deal with many problems involving dependent probabilities, whose solutions require the most delicate and careful applications of modern analysis. The rudimentary calculi which the outsiders find esthetically or philosophically pleasing are usually either insufferably awkward or completely insufficient for the needs of professionals. There is a strange situation, which one observer has facetiously described somewhat as follows: it is true with probability 1 that the technical workers in probability use the measure approach, but that the writers on "probability in general" descendants of Carlyle's professor, do not consider this approach worth much more than a passing remark.[6] The following pages outline how our previous treatment is generalized to deal with problems in which it is desirable to have the distribution of x_j vary with j (so that physically the experiments are no longer the same), and in which the x_j do not have to correspond to the results of independent experiments. Some attempt will also be made to show how the modern mathematical theory of real functions is applied to the probability calculus.

Let $x_j = x_j(\omega)$ be the jth coordinate of the point $\omega:(x_1, x_2, \cdots)$. Then as the sequence $\omega:(x_1, x_2, \cdots)$ varies, x_j does also: $x_j(\omega)$ is a function of ω. The functions $x_1(\omega), x_2(\omega), \cdots$ are functions defined on Ω, an abstract space on which a measure has been defined. Moreover Ω-measure has been defined in such a way that the Ω-set for which $x_j(\omega) < K$ (j, K fixed) is an Ω-set whose measure has been defined. (This set is composed of all sequences $(x_1, x_2, \cdots)$ whose jth coordinate is $<K$, and the measure is $F(K)$, using our last definition of Ω-measure.) In the terminology of measure theory, $x_j(\omega)$ is thus a measurable function. The study of the measure relations of Ω, and this is the whole of our probability calculus, can be considered, from this point of view, as the study of the properties of a sequence of measurable functions, one with very special properties, as we shall see, defined on some space. A measurable function defined on Ω is usually called a chance variable, in the theory of probability. (This terminology is somewhat dangerous, because it mixes Problems I and II.) The whole apparatus of modern real variable theory is applicable to these chance variables. Thus if $f(\omega)$ is a chance variable (measurable function of ω) (physically, a function of the observations), it is customary to define a number called its expectation. This number is simply the integral of $f(\omega)$, with respect to the given Ω-measure. The fact that the expectation of the sum of two chance variables is the sum of their expectations is simply the familiar theorem that the integral of the sum of two functions is the sum of their integrals. Let $S(j, K)$ be the Ω-set defined by the inequality $x_j < K$. Up to now we have supposed

[6] This analysis, like every other probability statement, is only an approximation to reality, but a fairly close one.

that the measure of $S(j, K)$ is independent of j, that is that the distribution of x_j is independent of j. We have also supposed that[7]

$$P\{S(1, K_1) \cdots S(n, K_n)\} = P\{S(1, K_1)\} \cdots P\{S(n, K_n)\} \tag{3}$$

for any positive integer n, and numbers $K_1, \cdots, K_n$. That is, we have supposed that $x_1(\omega)$, $x_2(\omega)$, $\cdots$ are mutually independent chance variables.[8] In fact probability measure on Ω has been defined just to make the foregoing two facts true. Mutual independence is a very strong hypothesis to impose on a sequence of functions. In many probability problems (Markoff chains for example), more general measures must be defined on Ω. The sequence $x_1(\omega)$, $x_2(\omega)$, $\cdots$ whose properties are those of Ω-measure, is then no longer a sequence of independent functions, and the distribution of x_j can vary with j.

At this level, the study becomes the study of any sequence of measurable functions, defined on some space of total measure 1. If f, g are given chance variables, they may turn out to be independent. In that case the theorem that the expectation of their product is the product of their expectations becomes, when translated into mathematical language, the familiar theorem that

$$\int\int f(x)g(y)\, dx\, dy = \int f(x)\, dx \int g(y)\, dy.$$

The mathematical theorems are not simply analogues of the probability theorems—they themselves are those theorems. When stated mathematically, the probability theorems need no proof: they need only recognition as standard results.

Empirical needs suggest that certain functions called conditional probability distributions, and conditional expectations, should be defined in a certain way. This is possible, as a formal matter,[9] and the theorems then proved about these functions gives them their usual meaning when translated into practical language. These functions are extremely useful tools in dealing with mutually dependent (that is not independent) chance variables.

The above approach is easily generalized to the stage needed in the study of Brownian movements or of time series, in which, instead of the proper initial

[7] $P\{S\}$ was defined as the measure of the Ω-set S.

[8] The n chance variables $f_1(\omega)$, $f_2(\omega)$, $\cdots$, $f_n(\omega)$ are said to be independent if for every set of n numbers $K_1, \cdots, K_n$, the following equality is true.

$$P\{f_j(\omega) < K_j, \quad j = 1, \cdots, n\} = \prod_j P\{f_j(\omega) < K_j\},$$

where $P\{\cdots\}$ denotes the Ω-measure of the Ω-set defined by the conditions in the braces. Thus in the example of a normal distribution in ν dimensions given above, $x_1, \cdots, x_\nu$ are independent functions on the space of ν dimensions, a fact which follows readily from the fact that the ν-dimensional density function is the product of ν functions of the separate variables.

[9] Cf. Kolmogoroff, loc. cit.

abstraction being a sequence $\{x_n\}$ of numbers, we have a one-parameter family $\{x_t\}$ (t takes on all real values). The number x_t may, for example, be thought of as the x-coordinate of a particle at time t. There is no difference in principle here: Ω is now the space of functions of t, instead of the space of sequences, that is functions of n. From the other point of view, instead of studying the properties of a sequence of measurable functions, it becomes necessary to study the properties of a one-parameter family of measurable functions.

DISCUSSION OF PAPERS ON PROBABILITY THEORY

By R. von Mises and J. L. Doob

1. **Comments by R. von Mises.** Professor Doob outlines a new theory of probability starting with the following three basic conceptions. First, he uses the notion of an infinite sequence of trials or better: of an infinite sequence of numbers $x_1, x_2, x_3, \cdots$ which can be considered as the outcomes of infinitely repeated uniform experiments. Second, he introduces (in his Theorem A) the limit of the relative frequency of a particular outcome α. Third, (in his Theorem B) the notion of place selection defined by a sequence of functions $f_n(x_1, x_2, \cdots x_{n-1})$ is employed. All these three concepts are completely strange to the so called classical theory as developed by Bernoulli, Laplace, Poisson, etc. They have been introduced and made the corner stone of probability theory in my papers published since 1919. I daresay that in no probability investigation before 1919 any of those notions even were mentioned.

This concerns what Professor Doob calls the Problem I or the purely mathematical aspect of the question. As to his Problem II or the relationship between the formal calculus and real facts Professor Doob stresses that the actual values for probabilities that enter as data into a particular argument have to be drawn from long, finite sequences of experiments. This is in complete accordance with the standpoint of my theory and in strict contradiction to the classical conception which knows only "a priori" probabilities determined by "equally likely cases."

In both theories, Professor Doob's and mine (not in the classical) a mathethematical model or picture is associated with a long sequence of uniform experiments. These models are different in both theories. My model (the "kollektiv") consists of one infinite sequence ω: $x_1, x_2, x_3, \cdots$ in which the limit of the relative frequency of each possible outcome α exists and is indifferent to a place selection; the value of this limit is called the probability of α.

On the other hand Professor Doob's model implies all logically possible sequences which form a space Ω and he shows that in this space a measure function can be introduced which fulfills the following conditions: (1) If m is a positive integer, the set of all sequences the mth element of which is α has a measure p_α independent of m; (2) the set of all sequences in which the relative frequency of α-results has either no limit or a limit different from p_α is zero; (3) if S is any place selection, the set of all sequences ω for which the relative frequency of α in $S(\omega)$ has either no limit or a limit different from p_α is likewise zero; this value p_α is called the probability of the outcome α. It then can be shown that a probability in this sense can be ascribed to certain events, i.e. to certain types of experiments which in some way are connected with the sequence of basic

experiments. E.g. if the original sequence consists of the single successive tossings of a die, the derived sequence may consist of pairs of tossings with the sum of the outcoming points as new value of α. The new probabilities p'_α are found as measures of certain sets in the original measure system established in Ω.

There is no doubt that the model used by Professor Doob for representing empirical sequences of uniform experiments is logically consistent. Its practical usefulness depends on how the usual problems of combining different kollektivs and so on can be settled within this scheme. This has to be shown in detail. It seems to me that my conception is simpler in its application and closer to reality, while his model may be considered more satisfactory from a logical standpoint since it avoids the difficulties connected with the concept of "all place selections." At any rate, however, there is no contradiction or irreconcilable contrast: both theories are essentially statistical or frequency theories, equally far from the classical conception based on "equally likely cases." In both theories probabilities are, of course, measures of sets.

2. **Comments by J. L. Doob.** It is perhaps unfortunate that Professor von Mises' treatment of probability problems, based on typical sequences ("collectives," "admissible numbers"), is commonly called the "frequency theory."[1] It is clear to any reader of our papers (identified as M and D below) that the idea of frequency, at least in the discussion of the relation of mathematics to practice, is no more fundamental to one approach than to the other. In one mathematical treatment frequency notions first appear in the theorems, whereas in the other they first appear in the axioms; but they appear in both. The principal objection the measure advocates have to the frequency approach is that it is awkward mathematically. Anyone who doubts this awkwardness need only examine various books published recently, using this approach, to see what a lot of fussy detail is involved merely in proving such elementary results as the Tchebycheff inequality or the Bernoulli theorem. One author considers it necessary to have his chance variables so restricted that if x is a chance variable, the event $x < k$ has a probability assigned to it only if k is not in some exceptional set, which may be infinite. To take another example, consider the coin tossing game discussed in both M and D, in which two out of three wins at tosses win a game. Apparently the probability analysis of this game is somewhat difficult in terms of the frequency theory. As the quite elementary treatment outlined in D shows, there is no difficulty involved, using the measure approach. The question is simple: a set of chance variables is given (corresponding to the original tosses); a new set is determined from them (corresponding to the grouping into games). Only elementary algebraic manipulation is required to verify that the new chance variables are mutually independent in the mathematical sense, (Cf. D), and have the same distribution, so the law of large numbers is applicable. Professor von Mises considers that the measure theory cannot handle this problem. I on the other hand consider that this problem exhibits the mathematical disadvantages of the frequency theory.

[1] This identifying name will be used below also.

The frequency theory reduces everything to the study of sequences of mutually independent chance variables, having a common distribution. "Probability theory is the study of the transformations of admissible numbers" writes Professor von Mises. This point of view is extremely narrow. Many problems of probability, say those involved in time series, can only be reduced in a most artificial way to the study of a sequence of mutually independent chance variables, and the actual study is not helped by this reduction, which is merely a tour de force.

It is claimed in M that the axioms of measure theory only describe the distribution within one collective (M, p. 00). This statement seems to mean that only the measure relations (using the notation of D) of the first coordinate function $x_1(\omega)$ can be discussed in the measure theory, that is only probabilities of the type: the probability that $x_1 < k$ (in the language of practice, "the probability that the result of the first experiment is less than k") are discussed. Actually, however, (Cf. D) the measure theory can discuss any number of experiments simultaneously, using the appropriate space Ω.

Many of the debates between the advocates of the various probability theories have been wasted, because some of the debaters talk mathematics, others physics. With this in mind, I should like to stress again[2] that (except for a few philosophically inclined Englishmen) everyone calculates probability numbers in the same way—a combination of reasoning based on experience and helped by theory, with examination of the experimental conditions and the results of trials. Frequency considerations necessarily play a large part. The fact that almost everyone calculates probability numbers in the same way does not alter the fact that one mathematical theory may be more useful or convenient than another in dealing with these probability numbers.

In closing, it seems proper to call attention to what the measure advocates consider the real services and contributions of the approach of Professor von Mises. Professor von Mises was the first to stress the importance of the second of two fundamental generalizations of experience in dealing with repeated mutually independent experiments of the same character: (1) the clustering of success ratios and (2) the fact that this clustering is unaffected by a system of rejection as described in M and D. These two generalizations of experience are certainly fundamental. The only point under discussion here is how such generalizations are to be put into a mathematical setting. The original such setting of Professor von Mises was criticized as not really mathematical. The setting now proposed by Copeland and others is criticized by the measure advocates as mathematically inflexible and clumsy. But it is significant that even in a treatment of the measure approach, as in D, it was felt essential to stress the mathematical interpretation of the two empirical generalizations of Professor von Mises. In the terminology of D, the measure advocates consider the contribution of Professor von Mises' approach to be a contribution to a solution of Problem II, not to Problem I, the mathematical problem.

[2] We are not talking mathematics now, but the application of mathematics.

CONTINUED FRACTIONS FOR THE INCOMPLETE BETA FUNCTION[1]

By Leo A. Aroian

Hunter College

1. Introduction. Existing literature on the problem of calculating the incomplete Beta function

$$B_x(p, q) = \int_0^x x^{p-1}(1-x)^{q-1}\,dx, \qquad 0 < x < 1, p > 0, q > 0, \tag{1.1}$$

and the levels of significance of Fisher's z [1] leave further work to be done. Müller's continued fraction and a new continued fraction are shown to possess complementary features covering the range of $B_x(p, q)$ for all values of x, p, q. Previous methods of computing $I_x(p, q) = B_x(p, q)/B(p, q)$ are given in [2], [5], [6], [8], [10], [13], [14], [15].

Müller's continued fraction is

$$I_x(p, q) = C\left[\frac{b_1}{1+}\,\frac{b_2}{1+}\,\frac{b_3}{1+}\,\frac{b_4}{1+}\cdots\right], \tag{1.2}$$

where

$$C = \frac{\Gamma(p+q)}{\Gamma(p+1)\Gamma(q)}\,x^p(1-x)^{q-1}, \qquad b_1 = 1, \qquad \mu_s = \frac{q-s}{p+s},$$

$$b_{2s} = -\frac{(p+s-1)(p+s)}{(p+2s-2)(p+2s-1)}\,\mu_s\,\frac{x}{1-x},$$

$$b_{2s+1} = \frac{s(p+q+s)}{(p+2s-1)(p+2s)}\,\frac{x}{1-x}.$$

A convergent infinite series $1 + \sum_{n=1}^{\infty} d_n x^n$ can be converted into an infinite continued fraction of the form $\frac{1}{1+}\,\frac{c_1 x}{1+}\,\frac{c_2 x}{1+}\cdots$ where [4], [9] p. 304,

$$c_1 = -\beta_1, \qquad c_2 = \frac{-\beta_2}{\beta_1},$$

$$c_{2s} = \frac{-\beta_{2s-3}\beta_{2s}}{\beta_{2s-2}\beta_{2s-1}}, \qquad c_{2s+1} = \frac{-\beta_{2s-2}\beta_{2s+1}}{\beta_{2s-1}\beta_{2s}}, \qquad s > 2 \tag{1.3}$$

---

[1] Presented at a meeting of the American Mathematical Society, October 28, 1939, New York City.

where

$$\beta_{2s} = \begin{vmatrix} 1 & d_1 & d_2 & \cdots & d_s \\ d_1 & d_2 & d_3 & \cdots & d_{s+1} \\ \vdots & & & & \vdots \\ d_s & d_{s+1} & d_{s+2} & \cdots & d_{2s} \end{vmatrix}, \qquad \beta_{2s+1} = \begin{vmatrix} d_1 & d_2 & d_3 & \cdots & d_{s+1} \\ d_2 & d_3 & d_4 & \cdots & d_{s+2} \\ \vdots & & & & \vdots \\ d_{s+1} & d_{s+2} & d_{s+3} & \cdots & d_{2s+1} \end{vmatrix},$$

$$\beta_{2s} \neq 0, \qquad \beta_{2s+1} \neq 0.$$

The infinite continued fraction found in this manner is called the corresponding continued fraction and the power series is said to be semi-normal if $\beta_{2s} \neq 0$, $\beta_{2s+1} \neq 0$.

2. A new continued fraction. Müller found his continued fraction by converting in the manner of the preceding paragraph

$$I_x(p, q) = \frac{\Gamma(p+q)x^p(1-x)^{q-1}}{\Gamma(p+1)\Gamma(q)} \cdot \left\{1 + \sum_{r=0}^{\infty} \frac{(q-1)(q-2)\cdots(q-r-1)}{(p+1)(p+2)\cdots(p+r+1)} \left(\frac{x}{1-x}\right)^{r+1}\right\}, \quad x < \tfrac{1}{2}. \tag{2.1}$$

We convert

$$I_x(p, q) = \frac{\Gamma(p+q)x^p(1-x)^{q}}{\Gamma(p+1)\Gamma(q)} \cdot \left\{1 + \sum_{r=0}^{\infty} \frac{(p+q)(p+q+1)\cdots(p+q+r)}{(p+1)(p+2)\cdots(p+r+1)} x^{r+1}\right\}, \quad 0 < x < 1. \tag{2.2}$$

Consequently

$$\beta_1 = \frac{p+q}{p+1}, \qquad \beta_2 = \frac{(p+q)(1-q)}{(p+1)^2(p+2)}, \cdots,$$

$$\beta_{2s+1} = \frac{(p+q)(p+q+1)\cdots(p+q+s-1)(p+q+s)}{(p+s+1)(p+s+2)\cdots(p+2s)(p+2s+1)} \beta_{2s},$$

$$\beta_{2s+2} = \frac{(1-q)(2-q)\cdots(s-q)(s+1-q)(s+1)!}{(p+1)(p+2)\cdots(p+2s+1)(p+2s+2)} \beta_{2s+1},$$

$$c_{2s+1} = -\frac{(p+s)(p+q+s)}{(p+2s)(p+2s+1)}, \qquad c_{2s} = \frac{s(q-s)}{(p+2s-1)(p+2s)},$$

and

$$I_x(p, q) = \frac{\Gamma(p+q)x^p(1-x)^q}{\Gamma(p+1)\Gamma(q)} \left\{\frac{1}{1+}\,\frac{C_1}{1+}\,\frac{C_2}{1+}\cdots\right\}, \tag{2.3}$$

where $C_s = c_s x$. By well known theorems due to Van Vleck [12] and Perron [9] p. 347 we find (1.2) converges for $-1 < x < \infty$, and (2.3) converges for $-\infty < x < 1$, and in the neighborhood of zero (2.2) equals (2.3). The region of equivalence of the series and the fraction may be extended by the following argument. Let the infinite series be terminated at some arbitrary point which gives the desired accuracy. Then the continued fraction of the corresponding type represents this finite series, is finite and gives the result within the desired accuracy. The new continued fraction may also be derived by use of the hypergeometric series [9] p. 348. A special case of (2.3) was given by Markoff [3], pp. 135–41, [11] pp. 53–55, who applied the result only to the binomial distribution. The associated continued fraction provides more rapid convergence than the corresponding continued fraction. The associated continued fraction is found by means of the hypergeometric series [9] p. 331, p. 348:

$$I_x(p,q) = \frac{\Gamma(p+q)x^p(1-x)^q}{\Gamma(p+1)\Gamma(q)}\left\{1 + \frac{k_1 x}{1 + l_1 x +}\,\frac{k_2 x^2}{1 + l_2 x +}\,\frac{|k_3 x^2}{1 + l_3 x +}\cdots\right\}$$

$$\tag{2.4} \begin{gathered} k_1 = \frac{p+q}{p+1}, \qquad l_1 = \frac{p+q+1}{p+2}, \\ k_{s+1} = \frac{s(s-q)(p+s)(p+q+s)}{(p+2s-1)(p+2s)^2(p+2s+1)}, \\ l_{s+1} = \frac{s(q-s)}{(p+2s)(p+2s+1)} - \frac{(p+s+1)(p+q+s+2)}{(p+2s+2)(p+2s+1)}, \qquad s \geqq 1. \end{gathered}$$

The disadvantage of (2.4) lies in the unwieldy form of computation. For properties of an associated continued fraction and the corresponding continued fraction in connection with convergence and the Taylor series reference is made to [9] p. 331 and pp. 302–303.

3. Properties of the corresponding continued fraction. Müller and Soper [5], [10], pointed out the inadvisability of integration through the mode $x = \dfrac{p-1}{p+q-2}$. In such cases we change $I_x(p, q)$ to $I_{1-x}(q, p)$. Müller has shown for his continued fraction that if we do not integrate through the mode (we assume this in the remainder of the paragraph) that convergents 2, 3, 6, 7, etc., will be greater than the true value and the remaining convergents will be less than the true value provided q is an integer. However, if q is not an integer, and is small ($q < 20$), it may happen that all convergents are above the true value. In such cases we may consider whether Müller's continued fraction may apply by estimating the remainder $I(p+s, q-s)$, after s reductions by parts [10].

For the new continued fraction also

$$|C_{2s}| = \left|\frac{s(q-s)}{(p+2s-1)(p+2s)}\cdot\frac{p-1}{p+q-2}\right| < 1,$$

$$|C_{2s+1}| = \left|\frac{(p+s)(p+q+s)(p-1)}{(p+2s)(p+2s+1)(p+q-2)}\right| < 1,$$

and $C_{2s+1} < 0$; $C_{2s} > 0$ unless $s > q$ when $C_{2s} < 0$. If $C_{2s} > 0$ then the convergents 2, 3, 6, 7, 10, 11, etc., will be above the true value and the other convergents will be below the true value. If $C_{2s} < 0$, then all convergents will be above the true value. In such cases, since a remainder for the continued fraction has not been found, it seems best to estimate $I_x(p + s, q - s)$ to obtain an idea of the error.

4. $I_x(p + s, q - s)$ and the equivalent continued fraction. Soper [10] has given the remainder after s reductions by raising p. This will furnish an upper bound of the error in the corresponding continued fraction after s convergents. The remainder, when $q - s$ is a negative integer, is approximately

$$\begin{aligned} &I_x(p + s, q - s) \\ &\quad = \frac{2 \sin (q - s)\pi\sqrt{\xi(\xi - 1)/2\pi(p + q)}}{\xi - x}\left\{\left(\frac{x}{\xi}\right)^{\xi}\left(\frac{1 - x}{\xi - 1}\right)^{1-\xi}\right\}^{p+q}, \qquad (4.1) \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{where } \xi = \frac{p + s}{p + q}. \end{aligned}$$

Another approach is to use the equivalent continued fraction, for $s - 1$ convergents of the equivalent continued fraction reproduces exactly s terms of the infinite series. The infinite series and the equivalent continued fraction for the infinite series are alike in all respects except form. By [9] p. 210, we find that the equivalent continued fraction for (2.3) is

$$W_1 = \frac{\gamma_1}{1 + \gamma_1 -}\,\frac{\gamma_2}{1 + \gamma_2 -}\,\frac{\gamma_3}{1 + \gamma_3 -}\,\frac{\gamma_4}{1 + \gamma_4 -}\cdots$$

where

$$\begin{aligned} &\gamma_1 = \frac{p + q}{p + 1}x, \qquad \gamma_2 = \frac{p + q + 1}{p + 2}x, \qquad \gamma_3 = \frac{p + q + 2}{p + 3}x, \cdots \qquad (4.2) \\ &\qquad\qquad \gamma_r = \frac{p + q + r - 1}{p + r}x, \end{aligned}$$

and

$$I_x(p, q) = \frac{\Gamma(p + q)x^p(1 - x)^q}{\Gamma(p + 1)\Gamma(q)}\,\frac{1}{1 - W_1}.$$

The equivalent continued fraction for Müller's continued fraction is given in [5], p. 292.

5. Numerical illustration. If A_v and B_v represent the numerator and the denominator of the v-th convergent of a continued fraction $\frac{a_1}{b_1 +}\,\frac{a_2}{b_2 +}\,\frac{a_3}{b_3 +}\,\frac{a_4}{b_4 +}\cdots$ then'

$$\begin{aligned} A_v &= b_vA_{v-1} + a_vA_{v-2} \qquad (5.1) \\ B_v &= b_vB_{v-1} + a_vB_{v-2}, \qquad v > 2. \end{aligned}$$

As an example we calculate $I_{.5}$ (2.5, 1.5), which could not be done by Müller's continued fraction.

Convergent	A	B	A/B
1	1	1	1
2	1	.42857143	2.3333333
3	1.015873016	.44444444	2.2857142
4	.66233767	.29292929	2.2610838
5	.64812966	.28671329	2.2605498
6	.46471308	.20559441	2.2603391
7	.441837914	.195475117	2.2603281
8	.33105492	.14646345	2.2603245
9	.30890766	.13666520	2.2603242
10	.23762461	.10512856	2.2603240
11	.21882154	.096809808	2.2603240

Using the value of the eleventh convergent we have, $I_{.5}$ (2.5, 1.5) = .28779339. Pearson [7], p. 30, gives .2877934 and Soper [10], p. 32 gives .28779341.

6. Discussion of the various methods. Müller's continued fraction encounters difficulties when q is small due to the possible divergence of the series on which it is based. In such cases the new continued fraction works admirably. Where "reduction by parts" [10] is advisable it would seem Müller's results will be better, while if "integration raising p" is preferable, then the new continued fraction would be necessary. The other methods suggested in the past lacked in some cases remainder terms; were in other cases too long; were feasible only in a limited range; or were only approximations. I am particularly indebted to Professor C. C. Craig under whose guidance this study was completed.

REFERENCES

[1] L. A. Aroian, "A study of R. A. Fisher's z Distribution and the related F distribution," unpublished manuscript.

[2] B. H. Camp, "Probability integrals for the point binomial," *Biometrika*, Vol. 16 (1924), pp. 163-171.

[3] A. A. Markoff, *Wahrscheinlichkeitsrechnung*, translated from the second Russian edition by Heinrich Liebmann. Leipzig: B. G. Teubner, 1912.

[4] T. Muir, "New general formulae for the transformation of infinite series into continued fractions," *Roy. Soc. Edinb. Trans.*, Vol. 27, p. 467.

[5] J. H. Müller, "On the application of continued fractions to the evaluation of certain integrals, with special reference to the incomplete Beta function," *Biometrika*, Vol. 22 (1930-31), pp. 284-297.

[6] K. Pearson (Editor), *Tables for Statisticians and Biometricians*, Part II, London: Biometric Laboratory, 1931, pp. ccxxv-ccxxvi.

[7] K. Pearson, *Tables of the Incomplete Beta-Function*, London: Biometrika Office, 1934.

[8] M. V. and K. Pearson, "On the numerical evaluation of high order Eulerian integrals," *Biometrika*, Vol. 27 (1935), pp. 409-423.

[9] O. PERRON, *Die Lehre von den Kettenbrüchen*, Leipzig: G. Teubner, 1913. (Pages refer to this edition.)
[10] H. E. SOPER, *Tracts for Computers No. 7, The Numerical Evaluation of the Incomplete Beta-Function*, London: Cambridge Univ. Press, 1921.
[11] J. V. USPENSKY, *Introduction to Mathematical Probability*, New York: McGraw-Hill Book Co., 1937.
[12] E. B. VAN VLECK, "On the convergence of algebraic continued fractions," *Am. Math. Soc. Trans.*, Vol. 5 (1904), pp. 253-262.
[13] J. WISHART, "Determination of $\int_0^\theta \cos^{n+1}\theta\, d\theta$ for large values of n, and its application to the probability integral of symmetrical frequency curves," *Biometrika*, Vol. 17 (1925), pp. 68-78.
[14] J. WISHART, "Further remarks on a previous paper," *Biometrika*, Vol. 17 (1925), pp. 469-471.
[15] J. WISHART, "On the approximate quadrature of certain skew curves," *Biometrika*, Vol. 19 (1927), pp. 1-38.

NOTES

This section is devoted to brief research and expository articles, notes on methodology and other short items.

NOTE ON THE DISTRIBUTION OF NON-CENTRAL t WITH AN APPLICATION

BY CECIL C. CRAIG

University of Michigan

If we adopt the notation recently used by N. L. Johnson and B. L. Welch [1], non-central t is defined by

$$t = \frac{z + \delta}{\sqrt{w}},$$

in which δ is a constant and z and w are independent variables, z being distributed normally about zero with unit variance and w being distributed as χ^2/f in which f is the number of degrees of freedom for χ^2.

In the paper referred to Johnson and Welch discuss some applications of non-central t and give suitable tables calculated from the probability integral of the distribution of this variable. Previously tables of this probability integral for the purpose of calculating the power of the t test had been given by J. Neyman [2] and Neyman and B. Tokarska [3].

It is the purpose of this note to call attention to a series expansion for the probability integral of non-central t which is simple in form and in most cases convenient for direct calculation. As an application of some intrinsic interest this series is used to compute in several numerical cases the power of a test proposed by E. J. G. Pitman [4] based on the randomization principle.

If for convenience we write,

$$\sqrt{w} = \psi, \ (0 \leq \psi \leq \infty),$$

we have for the joint distribution of $z + \delta$ and ψ,

$$df(z + \delta, \psi) = \frac{2(f/2)^{f/2}}{\sqrt{2\pi}\,\Gamma(f/2)} e^{-\frac{1}{2}(f\psi^2 + z^2)} \psi^{f-1}\, d\psi\, dz. \tag{1}$$

From this

$$\begin{aligned} df(t, \psi) &= \frac{2(f/2)^{f/2} e^{-\delta^2/2}}{\sqrt{2\pi}\,\Gamma(f/2)} e^{-\psi^2(f+t^2)/2 + \delta\psi t} \psi^f \, d\psi\, dt \\ &= \frac{2(f/2)^{f/2} e^{-\delta^2/2}}{\sqrt{2\pi}\,\Gamma(f/2)} e^{-\psi^2(f+t^2)/2} \sum_{r=0}^{\infty} \frac{(\delta t)^r}{r!} \psi^{f+r}\, d\psi\, dt, \end{aligned} \tag{2}$$

Now this series can be integrated term by term with respect to ψ over its range and we have,

$$(3) \qquad df(t) = \frac{(f/2)^{f/2} e^{-\delta^2/2}}{\sqrt{2\pi}\, \Gamma(f/2)} \sum_{r=0}^{\infty} \frac{\Gamma[\frac{1}{2}(f + r + 1)]}{r!} (\delta t)^r \left(\frac{2}{f + t^2}\right)^{\frac{1}{2}(f+r+1)} dt.$$

This series converges uniformly in any finite interval for t and it may be integrated term by term over the entire range for t or over any part of it. In particular, after some reduction, we get,

$$(4) \qquad \begin{aligned} P(0 \le t \le t_0 \,|\, f, \delta) &= \int_0^{t_0} df(t) \\ &= \frac{e^{-\delta^2/2}}{2} \sum_{r=0}^{\infty} \frac{(\delta^2/2)^{r/2}}{\Gamma(r/2 + 1)} I\left((r + 1)/2, f/2; \frac{t_0^2}{f + t_0^2}\right), \end{aligned}$$

in which $I\left((r + 1)/2, f/2; \dfrac{t_0^2}{f + t_0^2}\right)$ is the incomplete Beta-function in the notation of Karl Pearson. Often what is wanted is

$$(5) \qquad P(-t_0 \le t \le t_0) = e^{-\delta^2/2} \sum_{r=0}^{\infty} \frac{(\delta^2/2)^r}{r!} I\left((r + 1)/2, f/2; \frac{t_0^2}{f + t_0^2}\right).$$

Since the incomplete Beta-function is numerically less than unity it is seen that the series (4) or (5) converges rapidly for moderate values of δ such as will ordinarily occur in applications for small samples. The use of Pearson's tables of $I(p, q; x)$ will be convenient since interpolation will be required for only one of the three arguments.

As an application let us consider the test proposed by Pitman in the paper referred to above. Two independent samples, $x_1, x_2, \cdots, x_{N_1}$, and $y_1, y_2, \cdots, y_{N_2}$, have been drawn and it is desired in the absence of any information about the two populations from which the samples came to test the hypothesis that they have equal means. A test based on what may be termed the principle of randomization for this situation has been discussed by R. A. Fisher [5] and by E. S. Pearson [6]. It is as follows: Let the combined sample of $N_1 + N_2$ observations be separated into sets of N_1 observations, $u_1, u_2, \cdots, u_{N_1}$, and N_2 observations, $v_1, v_2, \cdots, v_{N_2}$, in all possible ways. For each such separation let the numerical difference of the means, $|\bar{u} - \bar{v}|$, be the spread. Then for a suitably chosen $\delta > 0$, we will reject the hypothesis of equal means if fewer than $100\alpha\%$ of the ${}_{N_1+N_2}C_{N_1}$ spreads exceed $|\bar{x} - \bar{y}|$, and otherwise not. It is clear that this test is fiducially valid independently of the populations actually sampled in the sense that if it be consistently followed for all such samples, the proportion of cases when the hypothesis is rejected when it is true will statistically approach α.

For all but very small samples it is very tedious to calculate the ${}_{N_1+N_2}C_{N_1}$

spreads and Pitman in his discussion shows that for quite moderate values of N_1 and N_2 the quantity,

$$w = \frac{\dfrac{N_1N_2}{(N_1+N_2)^2}(\bar{u}-\bar{v})^2}{\dfrac{\Sigma(x-\bar{x})^2+\Sigma(y-\bar{y})^2}{N_1+N_2}+\dfrac{N_1N_2}{(N_1+N_2)^2}(\bar{u}-\bar{v})^2} = \frac{\zeta^2}{\xi^2+\zeta^2}$$

has a distribution which in all but very exceptional cases is quite well approximated by a $B(\frac{1}{2}, \frac{1}{2}(N_1 + N_2 - 2))$-function. That is, the distribution of w for the ${}_{N_1+N_2}C_{N_1}$ spreads may for practical purposes be found from that of t, by a simple transformation, with $N_1 + N_2 - 2$ degrees of freedom.

It seems pertinent to make some inquiry into the power of such a test, that is, to make an attempt to learn something about the probability that such a test will fail to reject the hypothesis of equal means when it is in fact false. To do this it is now necessary to specify the populations which have actually been sampled. If we suppose that these populations are normal with equal variances but with unequal means which, with no loss of generality, may be taken to be μ and $-\mu$ respectively, the probability integral of the distribution of non-central t will give our answer.

If we set

$$\frac{t^2}{f+t^2} = \frac{\zeta^2}{\xi^2+\zeta^2},$$

we have

$$t = \sqrt{f}\,\zeta/\xi.$$

Also,

$$\xi^2 = \frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}\cdot\frac{N_1+N_2-2}{N_1+N_2} = \frac{f}{N_1+N_2}s^2,$$

in which s^2 is the usual estimate of the population variance σ^2 based on $f = N_1 + N_2 - 2$ degrees of freedom. Then

$$t = \frac{\bar{u}-\bar{v}}{s}\sqrt{\frac{N_1N_2}{N_1+N_2}}$$

and this is a central t if $\mu = -\mu = 0$, otherwise it is non-central. In the latter case we write (the test is made on $\bar{x} - \bar{y}$),

$$t = \frac{(\bar{x}-\mu)-(\bar{y}+\mu)+2\mu}{s}\sqrt{\frac{N_1N_2}{N_1+N_2}}$$

$$= \frac{z+\delta}{\psi},$$

in which,

$$z = \frac{(\bar{x} - \mu) - (\bar{y} + \mu)}{\sigma} \sqrt{\frac{N_1 N_2}{N_1 + N_2}},$$

$$\psi = s/\sigma,$$

and

$$\delta = \frac{2\mu}{\sigma} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}.$$

In applying Pitman's test for a given significance level α, one determines whether or not

$$P(w > w_0) \gtrless \alpha,$$

w_0 being the value of w calculated from the sample. This is equivalent to finding

$$P(t^2 > t_0^2),$$

for the proper f, in which

$$\frac{t_0^2}{f + t_0^2} = w_0$$

and this can be found from an ordinary table of the probability integral of the t-distribution.

For a numerical example let $N_1 = N_2 = 10$ so that $f = 18$. If we adopt a 5% significance level we have $t_0^2 = 2.101^2$ for the critical value. Let us suppose that $\mu/\sigma = 0.1$, and calculate the probability that the hypothesis that $\mu = 0$ will be rejected. We have $\delta = 0.1$ and

$$\frac{t_0^2}{f + t_0^2} = 0.1969.$$

Then

$$P(t^2 \leq t_0^2) = e^{-0.1}[I(0.5, 9; 0.1969) + 0.1\, I(1.5, 9; 0.1969) + \frac{0.01}{2!} I(2.5, 9; 0.1969) + \cdots]$$

$$= 0.9292.$$

Four terms of the series were enough to give this result. The probability of rejecting the hypothesis in this case is thus 0.0708.

The following tables show results for $\alpha = 0.05$ and 0.01, $\mu/\sigma = 0.1$, 0.2, and 0.5, and $N_1 = N_2 = 10$ and 20.

Values of $P(t^2 > t_0^2)$

$N_1 = N_2 = 10$

α \ μ/σ	0.1	0.2	0.5
0.05	0.0708	0.1355	0.5621
0.01	0.0165	0.0396	0.2940

$N_1 = N_2 = 20$

α \ μ/σ	0.1	0.2	0.5
0.05	0.0947	0.2345	0.8691
0.01	0.0251	0.0862	0.6730

In only one case was it necessary to calculate as many as ten terms of the corresponding series to obtain these values.

REFERENCES

[1] N. L. Johnson and B. L. Welch, "Applications of the non-central t-distribution," *Biometrika*, Vol. 31 (1940), pp. 362–389.
[2] J. Neyman, Statistical problems in agricultural experimentation," *Roy. Stat. Soc. Jour.*, Supplement, Vol. 2 (1935), pp. 127–136.
[3] J. Neyman and B. Tokarska, "Errors of the second kind in testing 'Student's' hypothesis," *Amer. Stat. Assn. Jour.*, Vol. 31 (1936), pp. 318–326.
[4] E. J. G. Pitman, "Significance tests which may be applied to samples from any populations," *Roy. Stat. Soc. Jour.*, Supplement, Vol. 4 (1937), pp. 119–130.
[5] R. A. Fisher, *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1935, Section 21.
[6] E. S. Pearson, "Some aspects of the problem of randomization," *Biometrika*, Vol. 29 (1937), pp. 53–64.

NOTE ON AN APPLICATION OF RUNS TO QUALITY CONTROL CHARTS

By Frederick Mosteller

Princeton University

In the application of statistical methods to quality control work, a customary procedure is to construct a control chart with control limits spaced about the mean such that under conditions of statistical control, or random sampling, the probability of an observation falling outside these limits is a given α (e.g., .05). The occurrence of a point outside these limits is taken as an indication of the presence of assignable causes of variation in the production line. Such a form

of chart has been found to be of particular value in the detection of the presence of assignable causes of variability in the quality of manufactured product. As recently pointed out, however, the statistician may not only help to detect the presence of assignable causes, but also help to discover the causes themselves in the course of further research and development. For this purpose, runs of different kinds and of different lengths have been found useful by industrial statisticians.[1] Quality control engineers have found, at least in research and development work, that a convenient indication of lack of control is the occurrence of long runs of observations whose values lie above or below that of the median of the sample. For example (as will be shown below), at least one succession of 9 or more observations above or below the median in a sample of 40 would be taken as evidence of lack of control at the .05 level; meaning that under conditions of control such a run would occur in approximately 5 per cent of the samples. Since this type of test has been found useful by quality control engineers, it is perhaps desirable to discuss the mathematical basis of such tests of control and provide a brief table for samples of various sizes at the significance levels .05 and .01.

The general distribution theory of runs of k kinds of elements, and in particular that of two kinds has been thoroughly investigated by A. M. Mood.[2] The purpose of this note is to give an application of the general method to quality control.

Let us consider a sample of size $2n$ drawn from a continuous distribution function $f(x)$. These are then arranged in the order in which they were drawn. We now separate the sample into two sets by considering the nth and $(n + 1)$st elements in order of magnitude, then if $x_i \leqq x_n$, x_i will be called an a, and if $x_i \geqq x_{n+1}$, x_i will be called a b. A run of a's will be defined as usual as a succession of a's terminated at each end by the occurrence of a b (with the obvious exceptions where the run includes the first or last element of the sample), and

[1] The use of "runs up" and "runs down" as well as runs above and below the arithmetic mean of a sample were briefly described in a paper by W. A. Shewhart, "Contribution of statistics to the science of engineering," before the Bicentennial Celebration of the University of Pennsylvania, September 17, 1940, to be published in the proceedings of that meeting. In a paper, "Mathematical statistics in mass production," presented before the American Mathematical Society in February, 1941, Shewhart discussed some of the advantages of using runs above and below the median and showed how by comparing runs of different types in a given problem it is often possible to fix rather definitely the source of trouble. The present note considers only the frequency of occurrence of "long" runs which are often used by research and development engineers to indicate the presence of assignable causes of variation. The occurrence of more than one such run in a given sequence, if distributed above and below the median value may also constitute valid evidence of the presence of more than one state of statistical control between which the phenomena may oscillate. The interpretation of long runs in this sense, however, is not considered in the present note.

[2] A. M. Mood, "The distribution theory of runs," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 367-392.

runs of b's are defined similarly. A run of a's may conveniently be called a run "below the median," and a run of b's a run "above the median."

We shall use Mood's notation throughout, i.e., r_{1i}, r_{2i}, $(i = 1, 2, \cdots, n)$ are the number of runs of a's and b's respectively of length i, and r_1, r_2 are the total number of runs of a's and b's; $\begin{bmatrix} m \\ m_i \end{bmatrix}$ will indicate a multinomial coefficient, and $\binom{n}{k}$ a binomial coefficient. Also we define

$$\begin{aligned} F(r_1, r_2) &= 0, \qquad |r_1 - r_2| > 1, \\ F(r_1, r_2) &= 1, \qquad |r_1 - r_2| = 1, \\ F(r_1, r_2) &= 2, \qquad |r_1 - r_2| = 0. \end{aligned}$$

Then the distribution of runs of a's for our case is

$$P(r_{1i}) = \frac{\begin{bmatrix} r_1 \\ r_{1i} \end{bmatrix} \binom{n+1}{r_1}}{\binom{2n}{n}}. \tag{1}$$

We would like to find the probability of at least one run of s or more a's. The coefficient of x^n in

$$[x + x^2 + \cdots + x^{s-1}]^{r_1}, \tag{2}$$

gives the number of ways of partitioning n elements into r_1 partitions such that no partition contains s or more elements, and none is void. Rewriting (2) we have

$$x^{r_1}[(1 - x^{s-1})]^{r_1} \sum_{t=0}^{\infty} \binom{r_1 - 1 + t}{r_1 - 1} x^t,$$

and the coefficient of x^n is just

$$\sum_{j=0}^{r_1} (-1)^j \binom{r_1}{j} \binom{n - j(s-1) - 1}{r_1 - 1}. \tag{3}$$

Then the probability that we desire, of getting at least one run of s or more a's is immediately given by

$$P(r_{1i} \geqq 1, i \geqq s)$$

$$= \frac{\sum_{r_1=1}^{n-s+1} \left[\binom{n-1}{r_1-1} - \sum_{j=0}^{r_1} (-1)^j \binom{r_1}{j} \binom{n - 1 - j(s-1)}{r_1 - 1} \right] \binom{n+1}{r_1}}{\binom{2n}{n}}.$$

Noting that when $j = 0$ in the inner summation we have just the total number of partitions, we get finally

$$(4)\qquad P(r_{1i} \geqq 1, i \geqq s) = \frac{\sum_{r_1=1}^{n-s+1} \binom{n+1}{r_1} \sum_{j=1}^{r_1} (-1)^{j+1} \binom{r_1}{j} \binom{n-1-j(s-1)}{r_1-1}}{\binom{2n}{n}}.$$

A similar result of course holds for the b's.

If we desire the probability of getting at least one run of s or more of either a's or b's, we compute the probability of getting no runs of this type and subtract from unity. Expression (3) multiplied by the total number of ways of getting no partitions of s or more b's for a given r_1, and then summed on r_1 gives exactly the number of ways of getting no runs of either a's or b's as great as s. This is

$$(5)\qquad \begin{aligned} A = \sum_{r_1 > n/s} &\left[\sum_{j=0}^{r_1} (-1)^j \binom{r_1}{j} \binom{n-1-j(s-1)}{r_1-1}\right] \\ &\cdot \left[\sum_{r_2=r_1-1}^{r_1+1} F(r_1, r_2) \sum_{i=0}^{r_2} (-1)^i \binom{r_2}{i} \binom{n-1-i(s-1)}{r_2-1}\right], \end{aligned}$$

and the probability desired is

$$(6)\qquad P(r_{1i} \geqq 1 \text{ or } r_{2i} \geqq 1 \text{ or both}; i \geqq s) = 1 - A \bigg/ \binom{2n}{n}.$$

In spite of the complex appearance of A, the sum can be rapidly calculated for any given s, n since the calculations for the sums on i and j need not be duplicated.

In the case of a quality control chart, we set a significance level α for a given n, this determines s the length of run of either type necessary for significance at the level chosen. Suppose we are interested only in runs occurring on *one side* of the median, say above, when $\alpha = .05$, $n = 20$ (i.e., sample size equal to 40). We determine the least value of s which will make the right hand side of equation (4) less than or equal to .05. It turns out that $s = 8$ for this case. This means that under conditions of statistical control, i.e., random sampling, one or more runs of length 8 or more, above the median will occur in approximately 5 per cent of samples of size 40. Naturally an identical result holds when we are considering only runs below the median.

On the other hand, if under the same conditions as given above ($n = 20$, $\alpha = .05$), we are using as our criterion of statistical control the occurrence of runs of length s or greater *either* above or below the median, we must determine the least value of s such that $1 - A/\binom{2n}{n} \leq .05$. This value turns out to be 9. In other words under conditions of statistical control at least one run of at least 9 will occur *either* above or below the median in less than 5 per cent of the cases on the average.

The following table gives smallest lengths of runs for .05 and .01 significance levels for samples of size 10, 20, 30, 40, 50.

	Runs on one side of median		Runs on either side of median	
$2n$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
10	5	—	5	—
20	7	8	7	8
30	8	9	8	9
40	8	9	9	10
50	8	10	10	11

If there is an odd number of individuals, say $2n + 1$, in the sample, we would choose the value of the median as the dividing line for our sample and treat the data as if there were only $2n$ cases, thus ignoring the median completely.

The following table[3] gives the probabilities of getting at least one run of s or more on *one* side, *either* side, and *each* side of the median for samples of size 10, 20, and 40.

	$2n = 10$			$2n = 20$			$2n = 40$		
Length of Run (s)	One Side	Either Side	Each Side	One Side	Either Side	Each Side	One Side	Either Side	Each Side
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	.976	.992	.960	1.000	1.000	1.000	1.000	1.000	1.000
3	.500	.667	.333	.870	.956	.784	.992	.999	.986
4	.143	.230	.056	.457	.640	.274	.799	.930	.668
5	.024	.040	.008	.178	.293	.064	.450	.650	.249
6				.060	.106	.013	.207	.346	.068
7				.017	.032	.002	.087	.158	.016
8				.004	.007	.000	.034	.065	.004
9				.001	.001	.000	.013	.025	.001
10				.000	.000	.000	.005	.009	.000
11							.002	.003	.000
12							.000	.001	.000
13							.000	.000	.000

One method of computing such a table is to use expression (4) to obtain the probabilities on one side, and to use (6) to get probabilities for either side. Then the probabilities for runs on each side may be computed by using the relationship

$$2P\text{ (one side)} - P\text{ (either side)} = P\text{ (each side)}.$$

[3]The author is indebted to Dr. P. S. Olmstead of the Bell Telephone Laboratories for kindly placing this table at his disposal. Dr. Olmstead has pointed out that these probabilities have been found very useful in research and development work.

TEST OF HOMOGENEITY FOR NORMAL POPULATIONS

By G. A. Baker

University of California

1. Introduction. In biological experiments it is often of interest to test whether or not all the subjects can be regarded as coming from the same normal population. If they have not come from the same normal population, usually the most plausible alternative is that the subjects have come from a population which is the combination of two or more normal populations combined in some proportions. The combination of normal populations is a "smooth" alternative to the hypothesis of a single normal population. Such non-homogeneous populations are not the only "smooth" alternatives, of course, but are included among the "smooth" alternatives. If there is reason to believe that the only deviation from a normal population is due to non-homogeneity, then the results of Professor Neyman in his paper [1] are available in studying this problem.

It is desirable not to make any hypotheses about the mean and standard deviation of the sampled population, but to base all computations and tests on the data contained in the sample. Such a viewpoint has been stressed in a previous paper [2] where it was shown that if the sampling is from a normal population, the probability of a deviation from the mean of a first sample of n measured in terms of the standard deviation of the sample is proportional to

$$\frac{dv}{\left(1 + \frac{v^2}{n+1}\right)^{n/2}}. \tag{1.1}$$

The result (1.1) and Neyman's results give rise to a test of homogeneity which is valid for "large" samples. Empirical results show that fairly conclusive evidence of non-homogeneity may be obtained with samples of 100. Samples of 50 or less may be suggestive but rarely decisive.

2. Development of Test. Suppose that a sample of $n + 1$ is drawn from a normal population. It can be regarded as being made up of a first sample of n and a second sample of one. The value of v corresponding to (1.1) can then be computed and its distribution function is (1.1). This partition, of course, can be made in $n + 1$ ways. That is, $n + 1$ values of v are determined from a random sample of $n + 1$ from the original parent. It is true that these values of v are not independent among themselves. The correlation between the values of v, to a first approximation at least, is of the order of $1/n$ and can be neglected if n is "large."

A suitable transformation as discussed in [3], [1] and elsewhere, transforms (1.1) into a rectangular distribution.

If the same computations are made when the sampled population is not

normal, then the resulting values obtained will not be rectangularly distributed. For instance, suppose that the sampled population is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \left(pe^{-\frac{1}{2}(x-m_1)^2/\sigma^2} + qe^{-\frac{1}{2}(x-m_2)^2/\sigma^2}\right) \tag{2.1}$$

we find that the distribution of v based on the first sample of 2 is a very complicated expression involving sums of exponentials and definite integrals of exponentials. To obtain a rectangular distribution if the sampled population is normal, the appropriate transformation to make is

$$\begin{aligned} v &= -\sqrt{3} \cot \pi u \\ dv &= \sqrt{3}\, \pi \csc^2 \pi u \, du. \end{aligned} \tag{2.2}$$

The resulting u-distribution for population (2.1) then is to be compared with the rectangular distribution in the interval from zero to one.

For "large" values of $n + 1$ and for symmetrical non-homogeneous populations composed of two normal components, the u-distribution will be symmetrical about $u = \frac{1}{2}$, less than one near the ends, greater than one for values of u moderately far from $\frac{1}{2}$ and less than one for values of u near $\frac{1}{2}$. A Neyman [1] Ψ_k^2 of order 4 will be necessary to detect a difference of this sort. If the non-homogeneous population of two components is skewed, the u-distribution will still show the same two-humped effect but may be skewed instead of symmetrical. A Neyman Ψ_k^2 of order 4 should still be computed, although Ψ_3^2 may be more significant.

The test then consists of:

(a) computing the $n + 1$ quantities

$$x_i' = \frac{x_i - \bar{x}}{\sqrt{n+1}\, s}, \qquad (i = 1, 2, 3, \cdots, n + 1) \tag{2.3}$$

where

$$\begin{aligned} n + 1 &= \text{number in the sample} \\ x_i &= \text{the observed values} \\ x_j &= \text{the observed values except } x_i \end{aligned}$$

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j, \qquad s^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

(b) making the transformation

$$u_i = \int_{-\infty}^{x_i'} \frac{y_0\, dx'}{(1 + x'^2)^{n/2}}, \qquad (i = 1, 2, 3, \cdots, n + 1)$$

(c) computing the first four Ψ_k^2's of Neyman's paper [1]

(d) comparing Ψ_k^2 with $\Psi_\epsilon^2(k)$ as found from the Incomplete Gamma Function Tables.

If n is large, say $n = 100$, then u is given approximately by the normal probability integral.

If n is small, the values of u are obtained from the Table 25 of Vol. 2 of Pearson's Tables.

Neyman's derivation assumes that $n + 1$ is large and that the u's are independent. In this case, if $n + 1$ is large, then the u's are nearly independent, and hence the test is valid. The same procedure can be applied for smaller samples. It can not be expected that small differences from normal in the sampled population can be detected with small samples. Empirical results indicate that samples of 100 are necessary for decisive results even when the differences of the sampled population from a normal homogeneous population are large. Samples of 50 may be suggestive and in very extreme cases might be decisive.

TABLE I

Empirical Sampling Results

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Ψ_k^2's for 51 from population A	.0001	.843	2.009	7.464
Ψ_k^2's for 101 from population A	.086	2.403	4.998	12.868
Ψ_k^2's for 101 from population B	.553	.927	7.472	7.485
Ψ_k^2's for 101 from normal	.017	.082	1.288	1.663
$\Psi^2_{(.05)}(k)$'s (Neyman [1])	3.842	5.992	7.815	9.488
$\Psi^2_{(.01)}(k)$'s (Neyman [1])	6.635	9.210	11.345	13.277

It is to be noted that the test makes no assumption about the parameters of the sampled population and does not group the data. The application of the test gives a unique result that does not depend on the judgment of the computer in any respect. In applying the usual chi-square test the computer must choose groupings. The choice of groupings as indicated in [5] may change the P-values to very different levels of significance.

3. Empirical results. Samples of 51 and 101 from population A, of 101 from population B, and of 101 from a normal population, were drawn by throwing dice. Populations A and B are given in [4]. Population A is symmetrical and distinctly bimodal. Population B is weakly bimodal and strongly skewed.

For samples from population A it is necessary to compute Ψ_4^2. For samples from population B it may be sufficient to compute Ψ_3^2. The non-homogeneity of the type of population A seems to be somewhat more detectable than of the type of population B. The sample from the normal parent shows close conformity with expectation.

In applying the proposed test for homogeneity the u-values for small independent sets of data can be combined to give a much larger number of u-values.

REFERENCES

[1] J. NEYMAN, "«Smooth Test» for goodness of fit," *Skandinavisk Aktuariedskrift*, (1937), pp. 149–199.

[2] G. A. BAKER, "The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample," *Annals of Math. Stat.*, Vol. 6 (1935), pp. 197–201.

[3] G. A. BAKER, "Transformations of bimodal distributions," *Annals of Math. Stat.*, Vol. 1 (1930), pp. 334–344.

[4] G. A. BAKER, "The relation between the means and variances, means squared and variance in samples from the combinations of normal populations," *Annals of Math. Stat.*, Vol. 2 (1931), pp. 333–354.

[5] G. A. BAKER, "The significance of the product-moment coefficient of correlation with special reference to the character of the marginal distributions," *Jour. Am. Stat. Assoc.*, Vol. 25 (1930), pp. 387–396.

A NOTE ON THE POWER OF THE SIGN TEST

By W. MAC STEWART

University of Wisconsin

1. Introduction. Let us consider a set of N non-zero differences, of which x are positive and $N - x$ are negative; and suppose that the hypothesis tested, H_0, implies, in independent sampling, that x will be distributed about an expected value of $N/2$ in accordance with the binomial $(\frac{1}{2} + \frac{1}{2})^N$. As a quick test of H_0, we may choose to test the hypothesis h_0 that x has the above probability distribution. Defining r to be the smaller of x and $N - x$, the test consists in rejecting h_0 and therefore H_0 whenever $r \leq r(\epsilon, N)$, where $r(\epsilon, N)$ is determined by N and the significance level ϵ.

2. Power of a test. In applying such a test it is of interest to know how frequently it will lead to a rejection of H_0 when H_0 is false and the situation H implies that the probability law of x is $(q + p)^N$, with $p \neq \frac{1}{2}$, thereby indicating an expectation of an unequal number of $+$ and $-$ differences. The probability of rejecting H_0 when H_1 implying $p = p_1$ is true, is termed the *power* of the test of H_0 relative to the alternative H_1.[1] Thus, from the point of view of experimental design the power (P) of the test of H_0 may be considered a function of the alternative hypothesis H_1, the significance level ϵ, and N. As such, the following observations may be noted:

1. The power P_2, for an assumed ϵ, N, and H_2 implying $p = p_2$ is greater than or equal to the power P_1 for ϵ, N and H_1 implying $p = p_1$ where $|p_2 - .50| > |p_1 - .50|$.

[1] For an extensive discussion of the power of a test, the reader is referred to J. Neyman and E. S. Pearson, *Statistical Research Memoirs*, Vol. 1 (1936), pp. 3–6.

2. The power P_2 for an assumed H_1, N, and ϵ_2, is greater than or equal to the power P_1 for H_1, N, and ϵ_1, where $\epsilon_2 > \epsilon_1$.

3. The power P_2 for an assumed H_1, ϵ, and N_2 is greater than or equal to the power P_1 for H_1, ϵ, and N_1 where $N_2 > N_1$.

Hence, to increase the power of the test of H_0 relative to a particular H_1, the methods implied in observations 2 and/or 3 may be employed. However, if any increase in an established ϵ is undesirable, the method implied in observation 3 is the alternative.

3. Explanation of table. In the interests of efficiency and economy, two questions then arise: (1) What is the minimum value of N, which, at the significance level ϵ, will give the test of H_0 a power $P > \beta$, relative to a particular alternative hypothesis H_1? (2) For this minimum value of N corresponding to ϵ, what is the maximum value of r? Stated in another manner, the questions are these: "What is the smallest number (min N) of paired samples that must be employed in conjunction with the Sign Test in order that the test of H_0, at the significance level ϵ, shall have a power $P > \beta$ relative to an alternative hypothesis H_1?" (2) If x of these paired samples give rise to a positive difference, and (min $N - x$) a negative difference, and if r be defined as the smaller of x and (min $N - x$); then, what is the maximum value that r may attain and still have the results, at the level ϵ, judged significant?

Table I provides the answers to these questions for the significance level $\epsilon \leqq .05$; and (1) for H_1 implies $p = p_1$ for values of p_1 from .60 to .95 (and thereby from .40 to .05) at intervals of .05; (2) for values of β from .05 to .95 at intervals of .05, and also for $\beta > .99$. For example, assume that a power $P > .80$ relative to the alternative hypothesis H_3 ($p_1 = .70$) is desired. In Table I, the entry appearing in the column headed H_3 ($p_1 = .70$), and in the row $P > .80$ is 49,17—indicating that 49 paired samples are required, of which 17 or less must be of one sign (+ or −) and hence 32 or more must be of the opposite sign in order that the results be significant at the .05 level.

Because of the discreteness of the binomial distribution, it is impossible to maintain the level of significance at .05 or even arbitrarily close to that figure and still hold to the criterion that N shall be at a minimum. For that reason, particularly when min N is small, results significant at .05 according to Table I may be significant at a level ϵ' where ϵ' is considerably less than .05. In general, however, and in particular when min N is large (greater than 50) both the quantities ($.05 - \epsilon'$) and ($P - \beta$) are small.

4. Illustrative example. Goulden[2] describes a simple experiment in identifying varieties of wheat. In this experiment, a wheat "expert" is presented paired grain samples of two particular varieties of wheat. The object of the

[2] C. H. Goulden, *Methods of Statistical Analysis*, John Wiley and Sons, New York, 1939, p. 2.

experiment is to test the ability of the expert to differentiate between the two varieties by arranging the pairs so that samples of one variety are on the left, say, and samples of the other variety are on the right.

In a problem of this type, it is desirable to have a sufficiently large number, N, of paired samples in order that the following conditions be fulfilled: (1) The probability that a person possessing no discriminating ability pass the test

TABLE I

Minimum number of paired samples and maximum values of related r

$H_0 \sim p_0 = .50$

(5% level of significance, i.e., $\epsilon \leq .05$)

(min N, max r)

Power	H_8 $p_1=.95$	H_7 $p_1=.90$	H_6 $p_1=.85$	H_5 $p_1=.80$	H_4 $p_1=.75$	H_3 $p_1=.70$	H_2 $p_1=.65$	H_1 $p_1=.60$
$0 < P \leq .05$	—	—	—	—	—	—	7,0	6,0
$P > .05$	—	—	—	—	—	7,0	6,0	9,1
$P > .10$	—	—	—	—	7,0	6,0	9,1	17,4
$P > .15$	—	—	—	8,0	6,0	9,1	12,2	25,7
$P > .20$	—	—	—	7,0	10,1	13,2	17,4	37,12
$P > .25$	—	—	8,0	6,0	14,2	12,2	23,6	44,15
$P > .30$	—	—	7,0	11,1	9,1	18,4	25,7	56,20
$P > .35$	—	—	6,0	10,1	12,2	17,4	30,9	65,24
$P > .40$	—	8,0	—	9,1	16,3	20,5	35,11	74,28
$P > .45$	—	7,0	11,1	—	15,3	26,7	42,14	89,35
$P > .50$	—	6,0	10,1	13,2	18,4	25,7	44,15	101,40
$P > .55$	—	—	9,1	12,2	17,4	30,9	51,18	112,45
$P > .60$	—	—	14,2	15,3	20,5	36,11	56,20	125,51
$P > .65$	7,0	11,1	13,2	19,4	23,6	35,11	63,23	143,59
$P > .70$	6,0	10,1	12,2	18,4	25,7	40,13	67,25	158,66
$P > .75$	—	9,1	16,3	17,4	28,8	44,15	79,30	175,74
$P > .80$	—	14,2	15,3	20,5	30,9	49,17	90,35	199,85
$P > .85$	11,1	12,2	18,4	25,7	35,11	56,20	101,40	227,98
$P > .90$	9,1	15,3	17,4	28,8	42,14	65,24	114,46	263,115
$P > .95$	12,2	17,4	23,6	35,11	49,17	79,30	143,59	327,145
$P > .99$	15,3	23,6	30,9	44,15	67,25	110,44	199,85	453,205

through sheer guesswork be less than ϵ; and (2) if past experience has proven that an expert *does* possess the ability to discriminate between the varieties to the extent of placing a proportion, p_1, of the pairs correctly in the long run, then the probability that he will pass the test be P.

Under these conditions, how large an N is required, and for that N, what is the maximum number of pairs that may be incorrectly placed without failing

the test? For alternative hypothesis H_4 ($p_1 = .75$), and for $P > .90$, referring to Table I, it is seen that 42 paired samples must be employed and not more than 14 may be placed incorrectly. Under the same alternative hypothesis, if it be required merely that $P > .50$ (i.e., an expert with an ability of .75 have better than an even chance of passing), then only 18 paired samples are necessary and not more than 4 may be arranged incorrectly.

Thus, before conducting an experiment in which the Sign Test is to be employed, if the experimenter first decides what power the test must have relative to a certain alternative hypothesis; then from the accompanying table he may learn the minimum number of paired samples that are necessary; and the related maximum value of r.

If this procedure is not followed, and an experimenter employs, say 6 paired samples, he may (as can be seen from the table) discover, to his dismay, that "experts" of ability .75 will be unrecognized more than 80% of the time.

MOMENTS OF THE RATIO OF THE MEAN SQUARE SUCCESSIVE DIFFERENCE TO THE MEAN SQUARE DIFFERENCE IN SAMPLES FROM A NORMAL UNIVERSE

By J. D. Williams

Phoenix, Arizona

The following result may have considerable application to trend analysis. The specific problem was proposed to me by R. H. Kent.

Consider a sample $0_n : X_1, X_2, \cdots, X_n$ from a normal population with zero mean and variance σ^2, the variates being arranged in temporal order. We seek the moments of the ratio of δ^2 to S^2, where

$$(n-1)\delta^2 = \sum_{j=1}^{n-1} (X_j - X_{j+1})^2 \tag{1}$$

and

$$nS^2 = \sum_{j=1}^{n} (X_j - \bar{X})^2. \tag{2}$$

Here $\bar{X}$ is the mean of the X_j. In order to simplify the algebra, we will work with quantities A and B defined by

$$\begin{aligned} 2\sigma^2 A &= (n-1)\delta^2, \\ 2\sigma^2 B &= nS^2. \end{aligned} \tag{3}$$

The characteristic function for the joint distribution of A and B is

$$\begin{aligned} \varphi(t_1, t_2) &= E(e^{At_1 + Bt_2}) \\ &= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \int\int_{-\infty}^{\infty} \cdots \int \exp\left(At_1 + Bt_2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n} X_j^2\right) \prod_{j=1}^{n} dX_j, \end{aligned} \tag{4}$$

where t_1 and t_2 are pure imaginaries. For the method of analysis which will be used here t_1 and t_2 will be considered as real variables. By straight forward methods we have

$$\varphi^{-2}(t_1, t_2) = \begin{vmatrix} a & b & d & \cdot & \cdot & \cdot & \cdot & \cdot & d \\ b & c & b & d & & & & & \cdot \\ d & b & c & b & d & & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & & d & b & c & b & d \\ \cdot & & & & & d & b & c & b \\ d & \cdot & \cdot & \cdot & \cdot & \cdot & d & b & a \end{vmatrix} \tag{5}$$

where the determinant is of nth order and its elements are

$$\begin{aligned} a &= 1 - t_1 - (n-1)T \\ b &= t_1 + T \\ c &= 1 - 2t_1 - (n-1)T \\ d &= T = t_2/n. \end{aligned} \tag{6}$$

It can be verified that the determinant has the value

$$\varphi^{-2}(t_1, t_2) = \sum_{j=0}^{n-1} \binom{2n-j-1}{j} (-t_1)^j (1-t_2)^{n-j-1}, \tag{7}$$

where the symbol $\binom{2n-j-1}{j}$ represents a binomial coefficient. From (7) we find the moments m_j of A/B as follows: Setting

$$t_2 = \sum_{k=1}^{j} t_{2k}, \tag{8}$$

we have

$$\begin{aligned} m_j &= \iint_{-\infty}^{0} \cdots \int \frac{\partial^j \varphi(t_1, t_2)}{\partial t_1^j}\bigg|_{t_1=0} \prod_{k=1}^{j} dt_{2k} \\ &= \frac{2^j \left[\dfrac{d^j}{dt_1^j} \varphi(t_1, 0)\right]_{t_1=0}}{(n-1)(n+1)\cdots(n+2j-3)}. \end{aligned} \tag{9}$$

The result is rather unexpected, for we have established that the moments of A/B are equal to the moments of A divided by the moments of B.

We find the following explicit values for the first few moments m_j:

$$
\begin{aligned}
&m_0 = 1\\
&m_1 = 2\\
(10)\quad &(n-1)(n+1)m_2 = 4(n^2+n-3)\\
&(n-1)(n+1)(n+3)m_3 = 8(n^3+6n^2+2n-21)\\
&(n-1)(n+1)(n+3)(n+5)m_4 = 16(n^4+14n^3+53n^2-8n-231).
\end{aligned}
$$

These are valid subject to the restriction $2n - 1 \geq j$, because in arriving at the explicit forms we have treated the binomial coefficient $\binom{k}{j}$ as if it were identically equal to $k(k-1) \cdots (k-j+1)/j!$.

From (10) it is easy to pass to the moments of $R = \delta^2/S^2$. For example, we find the mean value and variance of R to be

$$\frac{2n}{n-1}$$

and

$$\frac{4n^2(n-2)}{(n+1)(n-1)^2}$$

respectively.